Derivation and Validation of a Machine Learning Model for the Prevention of Unplanned Dialysis

by

Martin M. Klamrowski, B.Eng.

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Applied Science

in

Electrical and Computer Engineering
(Data Science Specialization)

Carleton University
Ottawa, Ontario

Abstract

Unplanned kidney dialysis is associated with higher morbidity and mortality rates among advanced chronic kidney disease patients. The incidence of unplanned dialysis can be attributed to myriad factors, but importantly, many of them are modifiable with appropriately timed intervention and treatment planning. A system tailored to the clinical question of the optimal dialysis preparation timeline could be crucial in mitigating the risk factors associated with initiating dialysis in an unplanned manner. Hereinafter, the development of clinical machine learning models for the prediction of kidney failure over short timeframes of 6 and 12 months is studied. The groundwork for the machine learning analysis is laid out, covering the characterization of The Ottawa Hospital's Multi Care Kidney Clinic dataset, the data processing, and a comparison of machine learning to traditional methods. We find that a data-driven approach proffers an opportunity to significantly reduce the burden of unplanned dialysis in advanced CKD centers.

Acknowledgements

My sincerest thanks go to my graduate advisors, Drs. Ran Klein, Jim Green, Greg Hundemer, and Ayub Akbari. I am indebted to them, as to the additional exceptional research collaborators who are building this project from the ground up – Drs. Chris McCudden and Babak Rashidi. Thank you for all that you do.

I am also thankful to Suzanne Jackson, Melanie Bujold, and Drs. Amber Molnar, Tim Ramsay, Fateme Rajabi, Cedric Edwards, Alissa Visram, and Matthew Oliver for the essential assistance and expertise provided along the way.

Thank you to the Canadian Institutes for Health Research for funding this project.

Thank you to my thesis committee, Drs. Ted Perkins, Jeff Gilchrist, and Sreeraman Rajan for taking the time to critically review and evaluate this thesis.

Table of Contents

List of Tables

Table of Figures

List of Abbreviations

uACR: urine albumin-to-creatinine ratio

AUC-ROC: area under the receiver operating characteristic curve

AUC-PR: area under the precision-recall curve

CI: confidence interval

CKD: chronic kidney disease

eGFR: estimated GFR

ESKD: end stage kidney disease

GFR: glomerular filtration rate

KDIGO: kidney disease improving global outcomes

KFRE: Kidney Failure Risk Equation

KGH: Kingston General Hospital

MCKC: Multi-Care Kidney Clinic

PrRe70: maximum precision at 70% recall

TOH: The Ottawa Hospital

UHN: University Health Network

Chapter 1: Introduction

1.1 Motivation

Chronic Kidney Disease (CKD) has emerged as a significant global burden, with its prevalence continuing to escalate at an alarming rate [1, 2]. The evolving pervasiveness of CKD underscores the necessity for more efficient management strategies of CKD patients, both to safeguard the sustainability of healthcare systems and to promote patient health and happiness. One of the clinical management challenges faced both by patients and providers alike is the incidence of unplanned dialysis starts, a term that refers to dialysis initiation in the inpatient hospital setting. In general, unplanned dialysis starts are associated with unfavorable outcomes, such as increased mortality and morbidity [3], thereby further complicating the patient's health status. The high incidence of unplanned dialysis (from 40-60% of all dialysis starts in this population [4, 5]) can be attributed to myriad factors [4-8]. Importantly, many of these risk factors are potentially modifiable, suggesting that the negative outcomes associated with unplanned dialysis starts could be mitigated through timely intervention. By addressing these modifiable risk factors, we could facilitate a more optimal transition to dialysis for CKD patients, thereby improving their prognosis and overall quality of life all while reducing the associated financial strain imposed on providers.

1.2 Problem Statement

The nature of this clinical challenge yields itself to predictive modeling, whereby a patient's immediate or future risk of kidney failure is estimated using statistical algorithms. While there have been numerous kidney failure risk prediction models developed and implemented over time [9-12], the rate of unplanned dialysis starts remains persistently

high. These models may be decoupled from the problem for two reasons. Firstly, they predict over longer timeframes of 2-5 years, when dialysis preparation should preferably occur 6-9 months prior to dialysis initiation [13]. Second, they are single-timepoint models, derived for more general CKD populations with better kidney function and a much lower risk of kidney failure. We hypothesize that a model tailored to predict the need for dialysis in advanced CKD settings should be dynamic (i.e., accounting for evolving clinical measures) and should predict at routine time intervals, and thus compliment the manner and style of patient monitoring in specialized CKD clinics. Altogether, this suggests that what may be lacking in advanced CKD practice is a model that is concretely tied to the described clinical question: must the patient be prepared for dialysis now, or not. That is, will the patient's kidneys fail in the next 6-9 months?

1.3 Contributions

This thesis contributes preliminary models and studies towards addressing the highlighted clinical question. The findings are presented herein and have been disseminated to the wider research community via one journal article and two conference presentations. Chronologically, they are:

- Klamrowski, M., et al., *POS-201 Machine Learning Prediction of Imminent Dialysis in Advanced CKD Patients.* Kidney International Reports, 2022. **7**(2): p. S86.
    - Presented as a poster at the World Congress of Nephrology in Kuala Lumpur (Virtual), and published as the conference abstract above, this study elicited the capabilities of a preliminary machine learning model for the prediction of kidney failure at timeframes at 3, 6, and 12 months. These analysis results were included as a component in a successful

CIHR Project Grant application (*Artificial Intelligence for the Prevention of Unplanned Dialysis*; 2022-2026; $195,000CAD; Pis, G. L. Hundemer, A. Akbari, C. R. McCudden, R. K. Klein, K. Thavorn).

- Klamrowski, M., et al., *Comparison of Machine Learning with Cox Regression Models for Kidney Failure Prediction among Patients with Advanced CKD.* CCMG-CSCC, 2023.
  - Poster presentation at the Canadian Society of Clinical Chemists 67th Annual Meeting in Winnipeg, Manitoba. Abstract and poster were prepared by M. M. Klamrowski. Work was presented by C. R. McCudden. This study demonstrated the potential for improved predictive performance using machine learning over Cox regression models for the prediction of kidney failure at short timeframes of 6 and 12 months.
- Klamrowski, M., et al., *Short Timeframe Prediction of Kidney Failure among Advanced CKD Patients.* Clinical Chemistry, 2023.
  - Published in Clinical Chemistry. This journal article introduced new models for use in advanced CKD contexts for the prediction of kidney failure at short timeframes of 6 and 12 months. A comparison was performed across several predictive model types prominent in the kidney failure prediction literature and using different sets of laboratory measurements: baseline alone versus baseline and follow-up visit data. The manuscript is included as **Chapter 4** of this thesis.

1.4 Thesis Structure

The outline for the chapters that follow is:

- *Chapter 2* lays the contextual and technical foundation for the analyses and studies that follow.
- *Chapter 3* describes the dataset, curated in a cohort of advanced CKD patients at the Ottawa Hospital's Multi-Care Kidney Clinic, on which all analyses were performed.
- *Chapter 4* is a manuscript entitled *Short Timeframe Prediction of Kidney Failure among Advanced CKD Patients* which studies the comparison of machine learning and Cox regression models. The manuscript has been published in Clinical Chemistry.
- *Chapter 5* is a second manuscript that aggregates all of the methodologies presented in *Chapter 3* for the derivation and external validation of a machine learning model for the prediction of kidney failure among advanced CKD patients. The manuscript is currently in preparation for submission.
- *Chapter 6* is a discussion of the field, including challenges and opportunities for future work. Finally, it draws conclusions from the work described herein.

Chapter 2: Background

This section outlines and details the requisite knowledge to understanding the main body of this work. **Section 2.1** is subdivided into three main subsections – providing the reader with some basic information on clinical kidney science (**Sections 2.1.1 – 2.1.4**), the management of CKD populations (**Section 2.1.5**), and individualized care pathways for CKD patients (**Section 2.1.6**). By no means are these sections comprehensive, given the constantly developing nature of CKD science and practice. Therefore, for further information a consultation of relevant texts such as *Chronic Renal Disease* by Kimmel and Rosenberg (2020) [14] and up-to-date guidelines such as Kidney Disease: Improving Global Outcomes (KDIGO) [15], is encouraged. **Section 2.2** presents the state-of-the-art methodology that is used in CKD progression prognostication using survival models and Cox regression. **Section 2.3** lays the informational foundation for the machine learning methodology that was used to derive the predictive models being proposed for use in advanced CKD clinics at The Ottawa Hospital and beyond.

2.1 Chronic Kidney Disease

Chronic kidney disease (CKD) is a progressive condition characterized by the gradual loss of kidney function over time, potentially leading to a range of symptoms and health complications that can become life threatening. CKD poses a considerable burden on individuals, families, and healthcare systems due to the constraints imposed on patient quality of life, increased healthcare costs, and the risk of progression to end-stage kidney disease (ESKD) requiring dialysis or transplantation.

Chronic kidney disease (CKD) affects millions of people worldwide [1]. In many low- and middle-income countries, CKD remains underdiagnosed and undertreated [1, 16]. In Canada, upwards of 4 million people have moderate, or advanced CKD progression [17, 18]. In the end stages of this disease, patients require life-saving kidney therapy through dialysis or kidney transplantation in order to sustain life. Many individuals with end-stage kidney disease (ESKD) experience a significant decline in their quality of life, as they must undergo costly and time-consuming treatments, all while being unable to work [17]. Alarmingly, the prevalence of CKD has been increasing over the past decades, driven in part by the rising prevalence of risk factors such as diabetes, hypertension, obesity, and an aging population [1, 2]. In Canada, the number of persons receiving life-saving kidney therapy is also steadily increasing, representing a significant financial and economic burden [19].

Understanding the complex interplay of factors contributing to the development and progression of CKD is essential for effective prevention, early detection, and successful application of management strategies. Extensive research has been conducted to elucidate the pathophysiological mechanisms underlying CKD, including the involvement of genetic predisposition [20], inflammation [21-23], oxidative stress [24], and metabolic abnormalities in general [25, 26]. Moreover, the identification of novel biomarkers such as cystatin C [27-29] and the advent of advanced imaging techniques [30, 31] further demonstrates the developing nature of the field of CKD research.

Despite these advancements, significant challenges persist. The heterogeneity and multifactorial nature of CKD etiology, and the interplay between genetic, environmental, and lifestyle factors make it a complex disease to manage. Moreover, the translation of research findings into clinical practice and the development of effective

interventions to halt disease progression or improve patient outcomes remain ongoing areas of investigation.

2.1.1 The Kidneys

The kidneys (illustrated in **Figure 2-1**) are organs in the urinary system of the human body, responsible for regulating the volume and chemical composition of fluids. Their primary role is to direct most water-soluble waste products, excess water, and toxins, from the blood into the urine which is eventually excreted from the body. Apart from filtration, the kidneys also help regulate electrolyte balance, blood pressure, and the production of red blood cells [32].

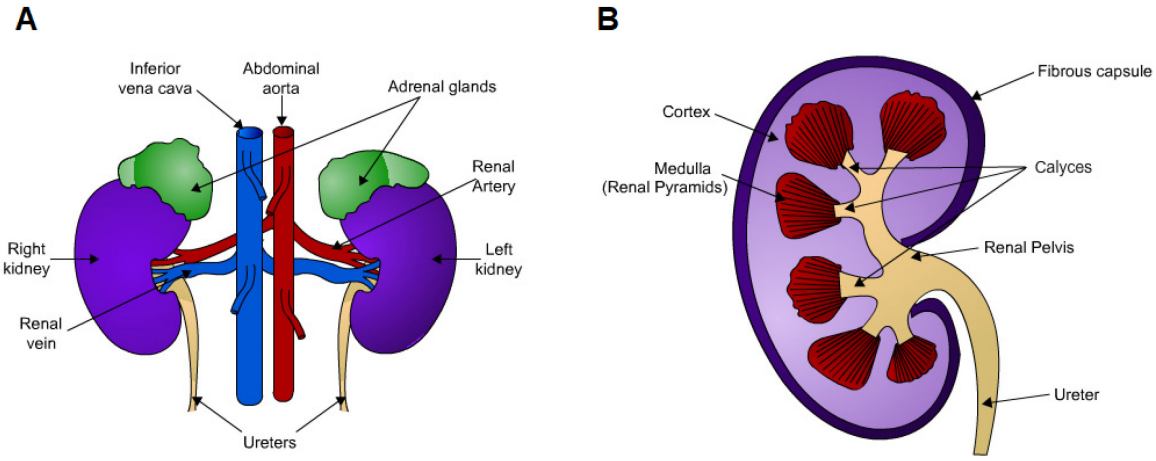


Figure 2-1: Principal illustration of the kidneys (**A**), and microstructures of the kidney seen on a cut surface (**B**). Duplicated from *The Urinary System* / University of the Highlands and Islands (CC0).

The two most commonly obtained measures of kidney function are the glomerular filtration rate (GFR), and albuminuria (albumin level in the urine). GFR – discussed in slightly more depth in **Section 2.1.2.1** – represents the rate at which blood is filtered by tiny blood vessels called glomeruli (singular: glomerulus). Each glomerulus inhabits a larger component called a nephron, which can be considered the functional unit of the whole kidney. Nephrons, generally speaking, are situated in the renal cortex and medulla

regions of the kidney (**Figure 2-1**). There are approximately one million nephrons in a healthy and functional kidney. As nephrons die, there are fewer functional glomeruli available to filter blood. As a result, the kidney's filtration rate (GFR) – its fundamental measure of function – decreases. Once sufficiently many nephrons die, waste products may remain in the blood, leading to the onset of CKD and its associated symptoms. Kidney dysfunction is also assessed using the protein, albumin. Albumin should normally be retained in the bloodstream by the kidneys. When kidney damage occurs, the filtration system becomes compromised, allowing albumin to leak into the urine. The presence of albumin in urine, known as albuminuria, is an early and important sign of kidney damage.

In summary, the kidneys play a crucial role in filtering waste and maintaining overall body function. When kidney function declines, measures such as GFR and urine albumin levels help assess the extent of damage. Substantial functional damage to the kidneys leads to CKD, and potentially kidney failure requiring dialysis or kidney transplantation (discussed in the sections to follow).

2.1.2 Definitions

Chronic kidney disease (CKD) is a clinical condition characterized by the progressive loss of kidney function over a period of months or years. The Kidney Disease: Improving Global Outcomes (KDIGO) defines CKD as albuminuria above 30 *mg/g*, an estimated glomerular filtration rate (eGFR; explained in **Section 2.1.2.1**) under 60 *mL/min/1.73m²*, or both, for three months or more, regardless of the underlying etiology [33]. CKD is further subcategorized into five stages based on the severity of kidney dysfunction, with stage 1 representing mild impairment and stage 5 indicating end-stage kidney disease (ESKD; explained in **Section 2.1.2.2**) requiring kidney replacement therapy. The complete KDIGO

classification system is depicted in **Figure 2-2**. The following sections describe in more detail the technical components that make up **Figure 2-2**.

				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 m²) 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			

Figure 2-2: KDIGO-specified stagewise classification of CKD. Green boxes denote low-risk prognosis of CKD; yellow denotes moderately increased risk; orange, high risk; red, very high risk. CKD is classified as persistent albuminuria ≥30 *mg/g* (A2) or eGFR <60 *mL/min/1.73m²* (G3), or both, for at least three months. Duplicated from the KDIGO Guidelines for Diabetes Management in Chronic Kidney Disease (CC0) [33].

2.1.2.1 Glomerular Filtration Rate (GFR)

Glomerular filtration rate (GFR) refers to the rate at which the kidneys filter waste products and excess fluid from the bloodstream, down into the urinary bladder where it is then expelled from the body (**Figure 2-1**). As such, GFR is the key indicator by which we assess kidney health and performance, and why its range is a useful scale for delineating CKD into the clinically meaningful stages introduced at the start of **Section 2.1.2**.

GFR is commonly expressed as milliliters per minute (*mL/min*) per 1.73 square meters (*mL/min/1.73m²*), where *mL/min* represents “flow rate”, and the latter (*1.73m²*)

represents a normalization to the average adult body surface area (BSA) [32, 34]. Given the function of the kidneys, this “flow rate” and its associated units are fairly intuitive and straight forward from a conceptual view. However, it may not be clear how this “flow rate” can be concretely measured in real kidneys. Clinically, “flow rate” is quantified with the aid of a physiological process called renal clearance. Renal clearance can be determined using plasma/serum samples, or a combination of urine and plasma/serum samples. Plasma and serum are two different components of blood that are available from a collected blood sample and separated out by centrifuging the sample. The choice between plasma and serum depends on the specific test being conducted and the requirements for the analysis. Renal clearance refers to the volume of plasma/serum from which a solute is cleared by the kidneys per unit of time [34]. More concretely, renal clearance is determined by comparing the rate at which a substance appears in the urine to its concentration in the blood plasma/serum. We may calculate the renal clearance of a plasma solute as

$$\frac{U_S \times V}{P_S},$$

where U_S represents the urine concentration of the solute, V is the urine flow rate (volume per time period), and P_S is the plasma concentration of the solute [32]. The ideal plasma solutes on which this clearance could be measured are inulin, iohexol, ^{51}Cr-EDTA, ^{99m}Tc-DTPA or ^{125}I-iothalamate [35].

It is important to note that the testing processes for the aforementioned markers are laborious and expensive, and thus are not clinically practical for routine follow-up of the patient. An alternative to these that has garnered widespread adoption is plasma/serum creatinine, given how closely its behavior mimics inulin. While not perfect, it has become the most commonly used marker for estimating renal clearance, and GFR.

Numerous equations to estimate the relationship between creatinine markers in the blood and measured GFR have been developed and validated in large multi-center cohort studies. These equations enable the estimation of GFR from easy-to-collect laboratory samples such as plasma/serum creatinine. Such estimating equations released for widespread use include the Modification of Diet in Renal Disease (MDRD) equation or the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation [36, 37]. Though new equations and studies are being implemented that do not include the use of a race modifier [27, 38].

In summary, it is important to understand that it is from creatinine, a surrogate for measured GFR, that a patient's kidney function is estimated, staged, and assessed. When GFR is estimated from creatinine, or similar surrogates, it is referred to as eGFR, where the "e" denotes "estimate". Clinical kidney science is evolving, and practices around creatinine collection and use may change. For example, currently cystatin-C is garnering increased attention due to its superior accuracy to serum creatinine in biomarking GFR [27-29, 35, 39]. However, its widespread adoption remains to be seen due to much higher costs and limited availability of instruments that would enable its routine implementation. While the physiological processes and laboratory practices just described are not inherently important to the understanding of this thesis, the uncertainty they introduce places important constraints on the interpretations of the findings herein. This uncertainty results from multiple factors including variations in sample collection and laboratory methodology and biological variations experienced by the patient throughout the day or between days.

2.1.2.2 End-Stage Kidney Disease (ESKD)

When the kidneys fail, they lose the ability to perform their vital functions adequately. This condition is known as end-stage kidney disease (ESKD) or kidney failure. As kidney function declines, waste products and toxins accumulate in the body, leading to symptoms such as fatigue, fluid retention, electrolyte imbalances, and high blood pressure. ESKD represents the final stage of CKD when the kidneys' functional capacity is severely compromised. At this stage, kidney replacement therapy, such as dialysis or kidney transplantation is required to sustain the patient's life. ESKD is associated with a higher risk of morbidity and mortality and requires ongoing medical management and support. Clinically, ESKD is classified based upon the patient's GFR (**Figure 2-2**): once a patient's GFR decreases to 15 *mL/min/1.73m²* or less, the patient is considered to have ESKD.

2.1.2.3 Dialysis and Transplantation

Two main modalities of dialysis exist: hemodialysis and peritoneal dialysis. Hemodialysis involves the extracorporeal removal of waste products and excess fluid through an artificial kidney (dialyzer) machine. It requires the pre-empted surgical creation of an access port to the blood called a fistula. In contrast, peritoneal dialysis uses the peritoneal membrane as a natural filter, where a dialysate fluid is introduced into the abdominal cavity, where the removal of waste occurs over several hours (e.g., during sleep), and then the fluid is removed. The choice of dialysis depends on various factors, including patient preference, clinical status, comorbidities, and the availability of resources.

Kidney transplantation, on the other hand, involves the surgical placement of a healthy donor kidney into a recipient with ESKD, offering the potential for improved quality of life and long-term survival. Kidney transplantation is considered the best treatment option for eligible CKD patients with ESKD. A successful kidney transplant offers the

potential for improved survival, better quality of life, and freedom from dialysis dependence. However, transplantation requires careful evaluation, immunosuppressive medications, and lifelong monitoring to prevent organ rejection, and is dependent on finding a suitable and willing donor.

2.1.3 Risk Factors

CKD development and progression are influenced by a range of risk factors, both modifiable and non-modifiable. Modifiable risk factors include diabetes mellitus, hypertension, obesity, smoking, sedentary lifestyle, excessive alcohol consumption, and certain medications. Non-modifiable risk factors encompass age, ethnicity (e.g., African descent), and certain genetic predispositions [40].

2.1.4 Biomarkers

Biomarkers play a crucial role in diagnosing and monitoring CKD. These measurable indicators in blood, urine, or other biological samples reflect the presence, severity, and progression of the disease. Examples of biomarkers used in CKD include serum creatinine, eGFR, albuminuria, cystatin-C, and various inflammatory and fibrotic markers. **Table 2-1** below lists the most common laboratory measurements utilized in the clinical management of CKD patients and what they are biomarkers for.

2.1.5 Management

Effective management of CKD aims to slow down the progression of kidney dysfunction, manage complications, and improve patient quality of life. The management strategies

encompass lifestyle modifications, pharmacological interventions, and, in the end stages of CKD, kidney replacement therapy.

Lifestyle modifications play a vital role in the management of CKD, particularly in the early stages. These include dietary adjustments, regular physical activity, smoking cessation, and weight management. Dietary recommendations often involve limiting sodium and phosphorus intake [41], moderating protein consumption, controlling fluid intake, and promoting a balanced diet rich in fruits, vegetables, and whole grains [42]. Additionally, patients may be advised to reduce the intake of potentially nephrotoxic substances, such as nonsteroidal anti-inflammatory drugs (NSAIDs) and certain herbal supplements [33, 43].

Table 2-1: Commonly collected laboratory measurements and the reasons for collection [14].

Laboratory Measurement	Biomarker	High Prevalence in Dataset (Yes/No)
Creatinine	Kidney Function	Yes
Blood Urea Nitrogen (BUN)	Kidney Function	Yes
Estimated Glomerular Filtration Rate (eGFR)	Kidney Function	Yes
Albumin	Nutritional Status	Yes
Hemoglobin	Anemia	Yes
Potassium	Electrolyte Balance	Yes
Phosphorus	Bone Health	Yes
Calcium	Bone Health	Yes
Parathyroid Hormone (PTH)	Bone Health	Yes
Sodium	Electrolyte Balance	No
Bicarbonate	Acid-Base Balance	Yes
Magnesium	Electrolyte Balance	No
Uric Acid	Kidney Function, Gout	No
Urine Protein	Kidney Damage, Proteinuria	Yes
Urine Albumin	Kidney Damage, Proteinuria	Yes
Urine Creatinine	Kidney Function, Proteinuria	Yes
Urine Red Blood Cells	Kidney Damage, Hematuria	No
Urine White Blood Cells	Kidney Infection, Inflammation	No

Regular physical activity has shown benefits in improving cardiovascular health, blood pressure control, insulin sensitivity, and overall well-being in CKD patients. However, exercise programs should be tailored to individual capabilities and may require adjustments based on the stage of CKD and presence of other comorbidities.

Pharmacological interventions are commonly employed to manage various aspects of CKD. Medications are prescribed based on the specific needs and comorbidities of individual patients. Some commonly used medications include:

- Angiotensin-converting enzyme inhibitors (ACE inhibitors) and angiotensin II receptor blockers (ARBs): These drugs are often prescribed to control hypertension and reduce proteinuria, thus slowing down the progression of CKD.
- Diuretics: Diuretics help manage fluid overload and edema, particularly in patients with CKD-related volume expansion.
- Phosphate binders: CKD patients often experience hyperphosphatemia, which can be managed with phosphate binders to reduce the risk of cardiovascular complications and mineral bone disorders.
- Erythropoiesis-stimulating agents (ESAs): ESAs stimulate red blood cell production and are used to manage anemia associated with CKD.
- Statins: Statin medications are prescribed to control dyslipidemia (unhealthy levels fats in the blood) and reduce the risk of cardiovascular events in CKD patients.

2.1.6 The Advanced CKD Clinic

In recent years, the establishment of specialized advanced chronic kidney disease (CKD) clinics has emerged as a valuable approach to optimize the management and care of patients with advanced stages of CKD. These clinics serve as dedicated centers where patients with CKD, particularly those approaching or already in ESKD, receive comprehensive and coordinated care. By providing specialized care, close monitoring, individualized treatment plans, patient education, coordination of renal replacement therapy, psychosocial support, and a commitment to research and innovation, the advanced CKD clinic plays a pivotal role in improving the outcomes and quality of life for patients with advanced stages of CKD. These clinics serve as a crucial link between primary care providers and nephrology specialists, promoting integrated and patient-centered care throughout the CKD journey.

2.1.6.1 Multidisciplinary Care Team

The management of CKD patients is a multidisciplinary effort involving nephrologists, dietitians, pharmacists, nurses, and other healthcare professionals. This collaborative approach ensures that patients receive holistic care that addresses the various aspects of their condition. The team members bring their expertise to bear on the diverse needs of CKD patients, including medical management, nutritional guidance, psychosocial support, and education about kidney replacement therapy options.

2.1.6.2 Disease Progression Monitoring

One of the primary roles of the advanced CKD clinic is to closely monitor disease progression in CKD patients. Regular monitoring of kidney function, blood pressure, electrolytes, and other relevant parameters is essential to guide treatment decisions and optimize patient outcomes. Through regular assessments of kidney function, such as

eGFR and proteinuria, the clinic can evaluate the rate of decline in kidney function and determine the appropriate timing for interventions or discussions regarding kidney replacement therapy. This proactive monitoring ideally allows for timely decision-making and planning, ensuring that patients are prepared for the transition to dialysis or kidney transplantation when necessary.

2.1.6.3 Patient Education

Education plays a pivotal role in the advanced CKD clinic, empowering patients to actively participate in their own care. The clinic provides comprehensive education on CKD, its progression, treatment options, and self-management strategies. Patients and their families are educated about the importance of adhering to medications, dietary restrictions, and fluid management. They are also informed about the potential complications of CKD and how to recognize and manage them. By empowering patients with knowledge, the clinic aims to enhance patient engagement, improve treatment adherence, and ultimately achieve better clinical outcomes [7, 44].

2.1.6.4 Psychosocial Support

CKD is associated with significant psychosocial challenges, including emotional distress, anxiety, depression, and financial burdens. The advanced CKD clinic recognizes the importance of addressing these aspects of patient care. Social workers and psychologists within the clinic provide counseling, emotional support, and assistance in navigating the financial aspects of CKD management. They also facilitate connections with support groups and community resources, which can offer additional support and a sense of belonging for patients and their families.

2.1.6.5 Individualized Treatment Plans

The advanced CKD clinic recognizes that CKD management requires an individualized approach, tailored to each patient's unique circumstances. The care team conducts comprehensive assessments, taking into account factors such as age, comorbidities, lifestyle, and personal preferences. Based on these assessments, personalized treatment plans are developed to optimize the management of CKD, including strategies for blood pressure control, glycemic control in diabetic patients, management of mineral and bone disorders, and prevention of cardiovascular complications [14].

It is common for patients in the advanced CKD clinic to forego kidney replacement therapy entirely, ultimately as a matter of personal choice. Such patients typically opt for *conservative care management*, where the goal is to maximize patient well-being without dialysis or kidney transplantation.

The clinic also provides guidance on lifestyle modifications, including dietary recommendations and physical activity, to support overall well-being and slow the progression of CKD.

2.1.6.6 Coordination of Kidney Replacement Therapy

For patients approaching the need for kidney replacement therapy, the advanced CKD clinic plays a crucial role in facilitating a smooth transition. The care team provides information about different modalities of dialysis (hemodialysis and peritoneal dialysis) and kidney transplantation, including their benefits, risks, and implications for lifestyle. They guide patients through the process of selecting the most suitable kidney replacement therapy option based on individual preferences, medical suitability, and availability of resources. The clinic also assists patients in accessing appropriate resources, such as referral to transplant centers or facilitating access to dialysis centers.

2.1.6.7 Outcomes

In practice, the type and incidence of possible patient outcomes varies from clinic to clinic and may be dependent on the admission criteria for the clinic as well as the availability of treatment resources. In The Ottawa Hospital's Multi-Care Kidney Clinic, the termination of a patient's follow-up typically coincides with either 1) dialysis initiation, 2) death from comorbidities, kidney failure, or other causes, or 3) dropout from the clinic and thus future observation. The analyses in this thesis include all such patients with these outcomes.

The initiation of dialysis may be subdivided based upon the treatment modality, and the setting in which dialysis was initiated. In this thesis, the focus is on unplanned dialysis starts which are associated with greater risk of unfavorable outcomes such as morbidity or subsequent death. Such patients are typically those beginning dialysis in an emergent manner that requires hospitalization and immediate and frequently unprepared dialysis treatment. This definition is the most complete definition of unplanned dialysis and is the one used throughout the thesis [5, 6]. I.e., those patients beginning dialysis in the inpatient setting, as opposed to the outpatient setting.

Unplanned dialysis represents a major burden on overall patient quality of life, as well as a major economic burden on healthcare providers. Patients typically have the option of either peritoneal dialysis or hemodialysis. Both options will significantly impact the patient's quality of life, but there are several tangible differences between these two modalities. Peritoneal dialysis can be considered preferable, as a less-imposing, safer, and cheaper alternative to hemodialysis [5, 6]. But unplanned dialysis patients are less likely to have undergone modality education or have had a peritoneal dialysis access port created, and usually initiate in-center hemodialysis. It follows that timely education and access creation in patients at risk would lead to less morbidity and mortality and lower hospital costs [5].

It should be noted that within a single clinic there is no single and precise definition for when a patient should be dialyzed because clinicians use a combination of symptoms and clinical data to drive these outcomes. In the context of predictive modeling, this represents a limitation because outcomes that cannot be concretely defined will be subject to bias and variability.

2.1.7 The Kidney Failure Risk Equation

A call-to-action to develop kidney disease progression risk scores, aided by an influx in rich data streams, has since dawned an era of rapid development and validation of predictive models for use in the clinical management of CKD patients [10-12, 39, 45-48]. In 2011, the Kidney Failure Risk Equation (KFRE) was developed [10], and soon after in 2016, it was validated internationally [49]. Since then, the KFRE has become the gold standard in predicting the risk of kidney failure at 2- and 5-year time horizons. The KFRE incorporates age, sex, eGFR, and urine albumin-to-creatinine ratio (uACR) to determine the longitudinal probability of kidney failure, making it easy to use in standard clinical practice. The rationale behind using KFRE lies in its ability to provide accurate risk estimation thereby facilitating appropriate interventions and treatment decisions. To this end, the KFRE has demonstrated excellent retrospective performance in diverse patient populations [9]. Clinical trials are currently underway to prospectively evaluate this model [50].

2.2 Survival Analysis

The study of patient CKD progression and risk factors frequently yields itself to survival analysis methodologies. Survival analysis is a traditional first choice for researchers interested in characterizing the time-to-event of an outcome (survival time) and the factors

that could be influencing different survival times among individual patients or groups. Survival analysis encompasses a vast toolbox of specialized statistical methods frequently leveraged by researchers in clinical trials, medical research, epidemiology, and social sciences. The following sections provide a primer on the survival analysis methodologies leveraged throughout this thesis. It is important to note that the theory that is elicited in the following sections is significantly tailored for relevance to the methods used in this thesis. The intricacies of survival and event history analysis, and even more broadly, counting processes, amalgamate into what is an expansive field of mathematical and statistical methods that are beyond the scope of this thesis. Most of the discussion centers around the Cox regression model. And so, discussions around the estimation and interpretation of hazard, for example, are strictly discussions of continuous-time hazard which may not generalize to a discrete-time case.

2.2.1 Survival Data

The key components of survival data are events and event times, obtained from the study of a multitude of individuals or observation instances. Survival datasets in medical research are often obtained from a cohort of individuals who are followed over time to track the occurrence of an event of interest. In the context of monitoring and managing advanced CKD patients, one instance in the survival dataset is represented by one patient. The event and associated event time for that patient would be an encoding of the patient's disease progression timeline. An illustrated example of this concept is presented in **Figure 2-3**.

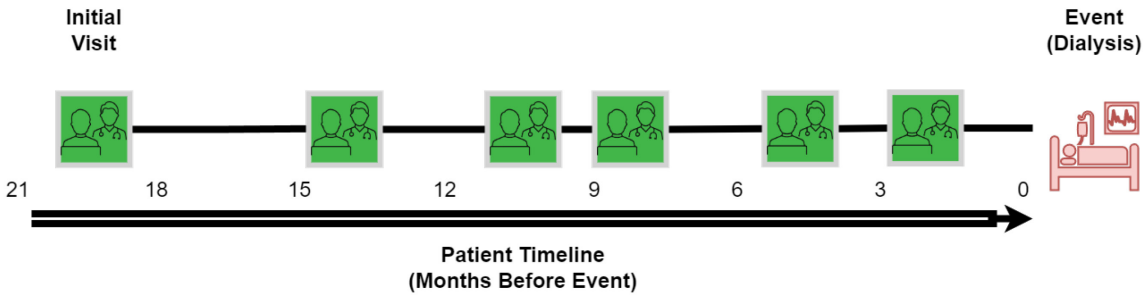


Figure 2-3: Abstract representation of an advanced CKD patient's timeline whilst under clinical management. Time zero is defined as the start of dialysis and during the months leading up to this event the patient had several visits to the CKD clinic (green boxes). The first visit to the CKD clinic is referred to as the initial visit and subsequent visits are referred to as follow-up visits.

In survival data, it is commonplace to have incomplete information on one or more individuals. Individual survival times are only definable insofar as the event time has been observed. When an event time is unknown, in that the event is assumed or known to have occurred but it is not known when, the observation is said to be censored. Censoring can occur for various reasons. For example, in advanced CKD studies, the most frequent reasons for the existence of censored patient data are 1) loss to follow-up, 2) the patient had not yet experienced the event, and 3) a competing event occurred before the event of interest.

It is important to distinguish between the different types of censoring [51]:

- *Right Censoring*: This is the most common type of censoring in most research studies and is the only type of censoring handled in this thesis. Right censoring occurs when an event has not yet occurred for an individual at the end of the study / data collection window.
- *Left Censoring*: Less common, this occurs when the exact event time is unknown, but it is known to have occurred before a certain time point. A common example of left censoring is when the event is assumed to have occurred before the start of the data collection window.

- *Interval Censoring*: Also less common, this occurs when the event is known to have occurred within a certain time interval.

To summarize, **Figure 2-4**, panels **A** and **B** visually present real survival data from the Ottawa Hospital's Multi-Care Kidney Clinic. Timelines for 10 patients split among each of the defined outcome groups are illustrated from their time of entry into the clinic until the end of their observed timeline. **Figure 2-4A** depicts the nature of the clinical survival data being dealt with here, whereby many patients were followed asynchronously over several years, with each timeline ending upon the observation of one or another outcome. Most survival data will traditionally be presented as seen in **Figure 2-4B**. **Figure 2-4B** demonstrates how most patient timelines typically follow the same pattern and differ only in their survival times and culminating events.

In advanced CKD, the event of interest is generally kidney failure, defined as the initiation of dialysis in an urgent, unplanned setting (UD) or had kidney replacement therapy (i.e., kidney transplantation or dialysis) in a planned setting (PD), in **Figure 2-4**). If the event of interest is not observed, it is said to be right-censored for that patient. The determination of what constitutes censoring is an important decision. This is especially true in medical research and advanced CKD given the high incidence of competing events (e.g., death), but holds true for any survival analysis study.

In closing, all of these data characteristics together constitute survival data and are most traditionally handled using survival analysis methods. With the methodology introduced in the following **Sections 2.2.2** and **2.2.3**, we can summarize survival data, model it, and use those artifacts to predict over future data.

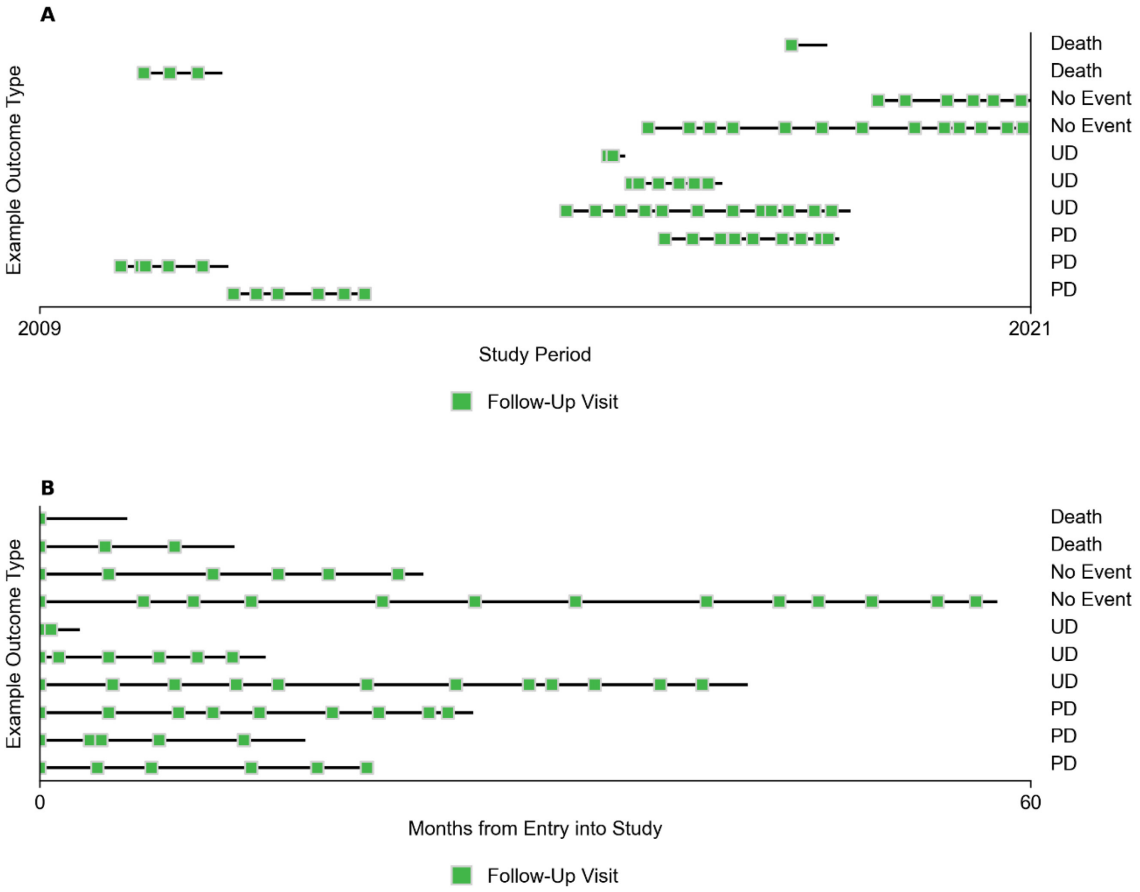


Figure 2-4: Illustration of patient survival data, with (**A**) each timeline distributed throughout the entire study period, and (**B**), each timeline's start left-aligned with 0 (i.e., time since entering the CKD clinic). A patient either died in pre-dialysis (Death), began dialysis in an urgent manner (UD), began dialysis in a planned manner (PD), or none of the above (No Event). The span of a patient's timeline is given by a black line. Points at which the patient was observed, and covariates recorded (follow-up visits) are marked in green.

2.2.2 Summarizing Survival Data

The analysis of time-to-event outcomes in survival data requires effective summarization techniques to capture underlying patterns and characteristics. This subsection focuses on the essential methods used to summarize survival data, estimate event times, and assess survival probabilities. In summarizing survival data, there are three functions of central interest. They are the survivor function, the hazard function, and the cumulative hazard function.

2.2.2.1 Survivor Function

The survivor function is a fundamental quantity in survival analysis and is estimated using nonparametric techniques. The survivor function, denoted generally by $S(t)$, provides the probability of survival beyond time t. Alternatively, it represents the probability of not experiencing the event of interest up to time t. This may be expressed as

$$S(t_j) = \mathrm{P}(T \geq t_j) \tag{1}$$

Where an individual's survival time can take any non-negative value T. Thus, $\mathrm{P}(T \geq t_j)$ is the cumulative probability of an event occurring beyond or at t_j. The subscript *j* indicates an index associated with a discrete time interval. Conversely, the failure time may be expressed as

$$F(t_j) = \mathrm{P}(T < t_j) \tag{2}$$

T is thus said to have a probability distribution with an underlying probability density function $f(t)$. $f(t)$ represents the density of probability at t and so can be integrated over to obtain the probability of survival (or failure) beyond (or before) a given time [51]. The probability of survival passed time t_j can therefore be expressed as

$$S(t_j) = \mathrm{P}(T \geq t_j) = 1 - F(t_j) = 1 - \int_0^{t_j} f(u)\mathrm{d}u \tag{3}$$

The survivor function ranges from 0 to 1 and is a non-increasing function over time. In other words, $S(t)$ decreases as t increases. It can be estimated using methods such as life tables, which are simply a summary of the survival times as demonstrated in **Figure 2-4B**, and the Kaplan-Meier estimator [52].

2.2.2.2 Hazard Function

Hazard is a quantity used to represent event occurrence. It represents the rate of the conditional probability of event occurrence. That is, the conditional probability of event occurrence per unit of time, which is a powerful and useful statistical measure given its specific and interpretable real-world meaning. Informally, the hazard function may be expressed as

$$h(t_j) = \lim_{\Delta t \to 0} \left\{ \frac{\mathrm{P}(t_j \leq T < t_j + \Delta t \mid T \geq t_j)}{\Delta t} \right\} \quad (4)$$

where for any t_j, we are interested in the probability per unit time that the survival time falls within $[t_j, t_j + \Delta t)$. In this case, the hazard function is expressing the expected number of events in a given time period (i.e., the expected event rate) [52]. The hazard function is a non-negative function and can vary over time. The hazard function defines the distribution of t, thereby determining both the density and survivor functions [53]. The hazard function can be estimated from a constructed grouped life table [52]. See **Section 2.2.5.1** for details on how the hazard function can be estimated using a Cox regression model..

2.2.2.3 Cumulative Hazard Function

The cumulative hazard function, denoted by $H(t)$, provides the cumulative risk of experiencing the event up to time t. It is the integral of the hazard function from time 0 to t (**Equation 5**). The cumulative hazard function is a non-decreasing function and can be estimated based on the estimated hazard function, specifically,

$$H(t_j) = \int_0^{t_j} h(u)du \quad (5)$$

Equation 5 above specifies a formal relationship between the hazard function and the cumulative hazard function. This is important because survival modeling (as will be discussed in **Section 2.2.5**) must often begin with the estimation of the hazard function. Then, as mentioned, the cumulative hazard function can be obtained using **Equation 5**. With $H(t_j)$ in hand, the survivor function can be obtained (calculus omitted [52]) from the following relation:

$$H(t_j) = -\ln S(t_j) = -\ln\left(1 - \int_0^{t_j} f(u)\mathrm{d}u\right) \tag{6}$$

2.2.3 Kaplan Meier Estimator

As previously mentioned, the Kaplan-Meier estimator provides a non-parametric method by which to estimate the survivor function of some survival data. It is given by [51]

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - e_j}{n_j} \ , \tag{7}$$

where n_j is the number of individuals that have survived up to time t_j, and e_j denotes the number of individuals experiencing the event in the j-th time interval. With the Kaplan-Meier estimator, each time interval is constructed such that a single event occurs at the start of the interval. **Figure 6A** illustrates the Kaplan-Meier estimate of the survivor function for the patient sample from **Section 2.2.1** and **Figure 2-4B**, where time zero represents the initial visit to the CKD clinic.

2.2.4 Comparing Survival Groups

2.2.4.1 Log-Rank Test

It is often of interest to compare survival experiences among different groups, such as treatment groups or patient subgroups. The log-rank test is a widely used statistical test for comparing survivor functions between two or more groups. It assesses whether the observed differences in survival curves are statistically significant. The log-rank test takes into account the observed event times and censoring information and compares the cumulative number of events in each group over time. To illustrate, consider the patient sample from **Section 2.2.1** and the associated Kaplan Meier estimate of the survivor function from the whole group (**Figure 2-5A, B**). In **Figure 2-5B**, the curve for the elevated creatinine group is consistently below (until 42 months) the curve for the reduced creatinine group, suggesting better survival outcomes for the reduced creatinine group. The log-rank test may be applied here to determine whether the observed difference in survival between the two groups is statistically significant. We obtain a p-value of 0.32, meaning the null hypothesis that the survival distributions are identical should not be rejected. While the provided example is likely to have been heavily influenced by sample size, it clearly illustrates the concept of survival differences among groups. Details of the log-rank test are not included here, for brevity. It suffices to understand that log-rank is a statistical test for comparing survival between different groups. To this end, log-rank will reappear in **Section 2.3** as the fundamental optimization criterion in random survival forests.

2.2.4.2 Hazard Ratio

The hazard ratio can be used to represent the ratio of the hazards between two comparison groups. While the log-rank assesses the statistical significance between

groups, the hazard ratio quantifies the extent of the observed difference between groups. For example, the hazard ratio between a group a and a group b is

$$\mathrm{HR} = \frac{h_a(t)}{h_b(t)} \,. \tag{8}$$

A hazard ratio greater than one indicates that group a is the higher-risk group, whereas a hazard ratio less than one indicates group b is the higher-risk group. Hazard ratios are used to make formal statements about risk among different comparison groups. The idea behind the hazard ratio (i.e., of relative risk between groups) also underpins the Cox regression model – to be discussed in the coming sections.

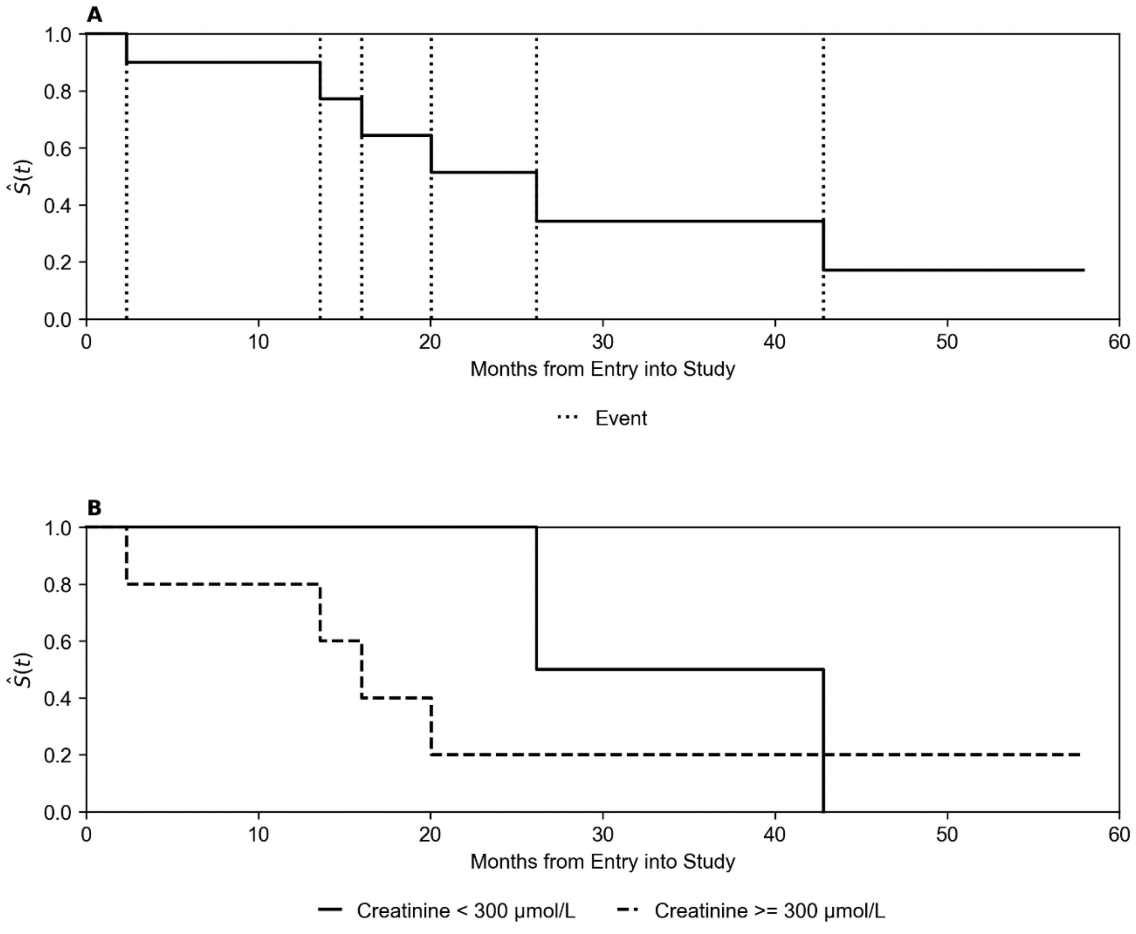


Figure 2-5: Kaplan Meier estimates of the survivor function for the patient sample introduced in **Section 2.2.1**. In (**A**), the survivor function for the entire group is estimated, and the event times are annotated. In (**B**), the

group is stratified into two subgroups based upon their initial creatinine measurement. Log-rank test p-value: 0.32.

2.2.5 Modeling Survival Data (Cox Regression)

While the aforementioned descriptors, when estimated, technically serve as descriptive models of survival data, they are limited. In survival analysis, it is frequently of interest to perform covariate-level analysis. I.e., how do unit changes in covariate values influence hazard and survival? The link and applicability to CKD is immediately evident, in that routinely collected clinical markers available from monitoring of CKD patients invite such analysis [6, 22, 54, 55]. For example, if a patient's hypertension is controlled, what impact may it have on survival? *Covariate*, *variable*, *feature*, and *predictor* are all terms that may be used interchangeably when referring to the input variables of a model. Herein, *covariate* is used.

This thesis focuses on only one of these models due to its widespread adoption in basic as well as clinical CKD science [9, 39, 49]: the Cox regression model (and an extension of the Cox regression model using time-varying covariates). Cox regression centers around the exploration of whether variations in covariates systematically influence event occurrence. Concretely, the Cox regression model stipulates that transformations in hazard scale linearly with covariates. This takes the form of a multiple linear regression:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{9}$$

where β_0 through β_p represent coefficients assigned to each of the p independent covariates given by $\boldsymbol{x}$. The linear component Y encodes this transformation in hazard using the concept of a hazard ratio. Here, the hazard ratio is expressing the hazard of an individual, i, relative to some reference, or baseline group (see below). To improve

distributional behavior, the hazard ratio is expressed on a log scale [52], yielding the final formulation

$$\log\left\{\frac{h(t_{ij})}{h_0(t_j)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\,. \tag{10}$$

Note the constant term, β_0, is absorbed by the baseline hazard, $h_0(t_j)$, which is discussed in the following passages [51].

Crucial to the Cox model is the definition of the reference/baseline group. I.e., $h_0(t_j)$ in **Equation 10**. The Cox model takes the hazard for this baseline group/individual to be an individual with all-zero covariates. Therefore, $\frac{h(t_{ij})}{h_0(t_j)}$ quantifies the hazard experienced by an individual, i, as characterized by that individual's covariate values, $\boldsymbol{X_i}$, relative to an individual with all-zero covariates. **Equation 10** is then easily transformed into the familiar Cox model expression of hazard,

$$h(t_{ij}) = h_0(t_j)e^{\beta_1 X_{1ij} + \cdots + \beta_p X_{pij}}\,, \tag{11}$$

or in terms of cumulative hazard,

$$H(t_{ij}) = H_0(t_j)e^{\beta_1 X_{1ij} + \cdots + \beta_p X_{pij}}\,. \tag{12}$$

Through the coefficient estimates, Cox models directly quantify changes in hazard per unit increase in covariate values. Given **Equation 11**, it can be seen that each β coefficient is in fact a logarithm of the hazard ratio (HR). As such, the Cox model's coefficients individually provide a direct measure of the relative risk being induced by that specific covariate. Together, they provide the overall relative risk of that individual's covariate set, or that individual's overall risk score.

Hazard ratios are reported in **Chapter 4** of this thesis for a number of laboratory measurements. The interpretation of a hazard ratio associated with a particular covariate in a fitted Cox model is straightforward. For example, if we have an estimated baseline

hazard of kidney failure of 0.04, and an individual's predicted or estimated hazard is 0.03, we say that the HR for the individual (individual vs. baseline) is HR = 0.03/0.04 = 0.75. With this result, and under the assumptions of a Cox model, we would make statements such as [56]:

- The individual has a 100% - 75% = 25% lower risk of kidney failure.
- The individual will have a (100% / 75%) – 100% = 33% increase in time before kidney failure.
- A group of many such individuals would experience the event at 0.75 times, or 75% the rate of the baseline group.

Despite their intended purpose, Cox models have garnered widespread usage as prediction models. In various disciplines of medical research, Cox models are used to longitudinally prognosticate the risk of an event occurring. In fact, the gold standard kidney failure risk prediction model is the KFRE, which is a Cox regression model. While the Cox regression model does not inherently provide an estimate of the baseline survivor function or the baseline cumulative hazard function, the recovery of these functions is possible using the estimated βs. Specifically, the obtained βs allow for an approximate estimate of the baseline cumulative hazard and baseline survivor function using the methods described in **Section 2.2.5.1**, and predicted survivor curves can then be obtained as

$$\tilde{S}(t_j) = \tilde{S}_0(t_j)^{r_i}, \tag{13}$$

where r_i is the individual's overall risk score ($e^{\beta_1 x_{1ij} + \cdots + \beta_p x_{pij}}$) [51]. An individual's predicted survivor curve represents a longitudinal continuous-time estimation of survival probability for that individual based upon the individual's covariate values. It is from this predicted survivor curve that models such as the KFRE produce predicted risk scores at the desired timeframe.

2.2.5.1 Estimating the Cox Model

Up to this point, methods for estimating the hazard function have only been alluded to. At a high-level, obtaining the fitted Cox model involves obtaining the coefficient estimates (β) through partial maximum likelihood estimation. From the coefficient estimates, the Breslow approximate estimate of the baseline cumulative hazard and baseline survivor function can be obtained [52]. With these quantities in hand the Cox model is specified, as per **Equation 11**, and estimates of survival can be computed using the equations in sections prior.

Obtaining the βs involves using the method of maximum partial likelihood. I.e., the betas are selected to be those that maximize the partial likelihood function, given by

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{r_i}{\sum_{l \in R(t_i)} r_l} \right\}^{\delta_i} . \tag{14}$$

Note that for computational efficiency, the logarithm of **Equation 14**, the partial log-likelihood, is what is most frequently implemented and optimized. In the context of the following discussion, the distinction is unimportant, and both the partial likelihood and the partial log-likelihood may and are used interchangeably. The maximum partial log-likelihood estimates for the β-parameters are obtained from the total product of individual relative risk contributions for each of the n individuals in the data (as shown in **Equation 14**). For the i-th individual's survival time, the contribution of the individual to the overall likelihood function is given by the individual's risk score at this time divided by the summed risk scores of all of the other individuals at risk (did not experience kidney failure yet) and uncensored ($R(t_i)$) up to that time. δ_i is an event indicator, defined for the i-th individual as

$$\delta_i = \begin{cases} 0, & \text{censored} \\ 1, & \text{otherwise} \end{cases} .$$

This means that censored individuals only contribute indirectly to parameter estimates via the summation over the remaining risk set, $R(t_i)$, in the denominator – assuming those individuals have a survival time greater than 0. The treatment of censored individuals for the purposes of survival analysis modeling in advanced CKD is under continued evaluation for this reason given the potential for biased parameter estimates [47].

Parameter estimates are obtained through an iterative procedure such as the Newton-Raphson procedure or gradient descent. The Newton-Raphson is commonly implemented in survival analysis software and so it is the procedure used in this thesis. The estimated β-parameters at the $(s+1)$-th iteration under Newton-Raphson are given by

$$\boldsymbol{\beta}_{s+1} = \boldsymbol{\beta}_s + \boldsymbol{I}^{-1}(\boldsymbol{\beta}_s)\boldsymbol{u}(\boldsymbol{\beta}_s)\,,$$

where $\boldsymbol{I}^{-1}$ is the inverse of the Fisher information matrix (which in this case is also the Hessian matrix taken at the negative log-likelihood function), and $\boldsymbol{u}$ is a vector of first derivatives of the log-likelihood function with respect to each β-parameter. Iteration starts at $\boldsymbol{\beta} = [0,0,\ldots 0]$, and terminates once the change in the log-likelihood is sufficiently small or the largest change across the β-parameters is sufficiently small. Once the Cox model parameters are obtained, the cumulative baseline hazard function and baseline survivor function can be recovered.

As an added note, **Equation 14** does not accommodate tied survival times – a common occurrence in studies with a sufficiently large sample size [51]. This places ambiguity on which individuals to include in the risk set underlying the denominator when computing the partial likelihood contribution for each of the tied individuals. A common and simple approximation is the Breslow estimate, which simply ignores the ties and

includes a term for each of the individuals with tied survival times. In each of these terms, all of the other tied survival times are included in the risk set in the denominator.

2.2.5.2 Time Varying Covariates

Up to now, the assumption has been that the Cox model incorporates covariates whose values remain constant over time. That is, the covariate value does not change with time. In clinical data, this is often not the case, and a patient may have a covariate measured at multiple timepoints throughout the study. The Cox model can accommodate the analysis of these types of covariates. **Equation 11** is already formulated to accommodate time varying covariates. Covariates in $\boldsymbol{x}$ will simply take on the time-updated covariate value in the j-th time period.

Time varying covariates do introduce some complexities in obtaining predicted survivor curves compared to the non-time varying Cox model. Under time varying covariates, **Equation 13** no longer holds. In lieu, the relation

$$\widetilde{P}_i(t, t+h) = \exp\left[-\left\{\widetilde{H}_0(t+h) - \widetilde{H}_0(t)\right\} * r_i\right] , \qquad (15)$$

given by Altman and De Stavola allows for the calculation of an approximate conditional probability of survival through an interval $t, t+h$, and thus can be used to prognosticate an individual's future survival probability [57].

Estimation of the parameters under time-varying covariates is mostly the same. I.e., **Equation 14** is used, but instead of assuming time-constant covariate values, the most recently available covariate values are used to compute the respective risk scores at each event time.

2.3 Machine Learning

In subsequent chapters, several machine learning algorithms were applied for the prediction of kidney failure over short timeframes. This section lays out the requisite machine learning knowledge for understanding the findings that are presented in later chapters.

Machine learning is frequently defined as a “subset of artificial intelligence (AI)”, where specific types of algorithms are exploited for their ability to automatically extract patterns from raw data without being provided any explicit instructions or rules [58]. One could argue, with some pedantry, that the “machine learning is a subset of AI” buzz phrase is somewhat ambiguous and inconsequential as a classification. Specifically, machine learning is the statistical estimation of functions. It is computational statistics grounded in a set of unique machine learning practices and paradigms. While this means that simple statistical models such as the Cox regression technically constitute machine learning, the term *machine learning* is more typically invoked with the application of algorithms that go beyond pure optimization (and can learn) to large datasets that contain many covariates. For example, the individual decision trees in random decision forests (introduced in **Section 2.3.2**) are simple models that directly optimize an objective function with respect to the data. Aggregating many such trees together into a random forest does not include new optimization criterion or change the objective functions of the individual decision trees. Yet, the learning ability is greatly improved by this aggregation, and deeper patterns can be extracted [59]. In contrast to models like Cox regression, much of the *learning* in machine learning occurs in this indirect manner.

Machine learning algorithms can be broadly categorized into supervised, unsupervised, and reinforcement learning algorithms. While each has the potential to be applied to clinical problems of this manner, only the first is used and presented in this

thesis. Thus, the goal in supervised learning is to build a model able to take input features (covariates) and generate the desired response. The term *supervised* stems from the paradigm through which this is done. When using a supervised learning algorithm to build a model, there is a prerequisite that the desired response for each instance in the set of input data used to train the model is known. This information is explicitly provided when estimating the parameters of the model.

2.3.1 Evaluation Metrics

Several evaluation metrics are pertinent to the studies and analyses contained within this thesis. Clinical prediction models are typically judged over two fundamental axes: discrimination and calibration. A good model is said to be well-calibrated and good at discriminating between data observations.

Discrimination refers to a model's ability to discriminate between observation types (e.g., event or no event, diseases or no disease, patient cohorts, disease class). In the context of this work, a model should predict higher risk of kidney failure for someone about to experience kidney failure than for a person with stable CKD. A discrimination metric summarizes over all of the test instances how a model’s prediction agrees with the actual observed outcomes. That is, a discrimination metric would typically report some aggregate measure of how good a model is at sorting patients in relation to one another or how accurately it predicts the actual outcomes. An example of a discrimination metric that is frequently reported in CKD progression studies is the concordance index. The concordance index reports the rank correlation between predicted risk and actual survival times. I.e., how good is a model at sorting pairs of observations.

On the other hand, calibration measures the alignment between predicted risk probabilities and the actual probabilities of the outcomes. A well-calibrated model

produces predicted probabilities that accurately reflect the true likelihood of an event occurring. This renders model predictions interpretable. For example, for a group of patients with predicted probability of kidney failure in the range of 50 to 60%, the actual incidence of kidney failure should be in the same range for a well-calibrated model. Calibration is traditionally reported using calibration curves plotting the predicted risk vs. the observed risk stratified by risk quintile or decile.

2.3.1.1 Classification

The discrimination and calibration metrics discussed in the prior section can be categorized as describing the goodness of fit of a model. However, they are limited in their ability to inform on the actual clinical utility of a model [60]. Classification metrics may be of benefit in this regard. A model being assessed on a classification problem is tasked with correctly assigning a class label (e.g., prepare for dialysis or not) to data instances. To evaluate models on this task, a confusion table can be constructed counting the combination of predicted and actual class-memberships for all subjects. **Table 2-2** shows a confusion matrix for a binary (two possible) class scenario. From this table, it may be understood that for a classifier correctly predicting and assigning a positive label to an instance of the positive class is a true positive (TP) prediction. To predict negative for that same instance would yield a false negative (FN). Likewise, for a negative outcome group, true negative and false positive predictions must be tabulated. In the context of a prediction model advising on a clinical decision, true positives and true negatives represent correct decisions, while false positives and false negatives represent incorrect guidance. From these base quantities, a number of important classification metrics may be derived. Each of these metrics directly provide a quantitative measure to help in understanding a model's effectiveness in predicting one or the other class label and the

types of errors the model is producing. The combination of reported metrics depends on the intended application. For example, if the minimization of false positives is paramount, the precision metric will likely be used as the primary evaluative marker for the model. Regardless, a confusion matrix must be summarized with at least two parameters to reflect the trade-offs in selecting an outcome probability cut-off when devising a classifier.

Table 2-2: The confusion table and its metrics.

Confusion Table

Actual \ *Predicted*	**False**	**True**
False	TN	FP
True	FN	TP

Metrics

$$\text{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

$$\text{Sensitivity} = \text{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

$$\text{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

$$\text{F1} = 2\frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Abbreviations: TN, true negative; FP, false positive; FN, false negative; TP, true positive.

2.3.1.2 Regression and Brier Score Metrics

A number of regression metrics are pertinent to the understanding of this thesis. Regression metrics are required when predicted values are being compared with a continuous target. The metrics used are the mean of the absolute errors (MAE), and the root of the mean of the squared errors (RMSE). Respectively, they are defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - \bar{y}_i|$$

and

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}$$

These metrics quantify the numerical accuracy of predicted values ($\hat{y}_i$) with each true value ($\bar{y}_i$). They are used to assess the performance of the imputation methods explored in **Section 3.4.1**.

Similar to the RMSE, but intended for a binary target, is the Brier Score. The Brier Score amounts to the mean squared error (RMSE^2) between a model's probabilistic predictions ($\hat{y}_i$) and the true binary outcomes ($\bar{y}_i$), thereby making it a measure of model calibration:

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$$

2.3.2 Random Decision Forests

The random decision forest algorithm, or a random forest, is a bagged ensemble method for supervised machine learning [61]. Random forests *bag* an ensemble of k decision trees. Meaning, these k decision trees are trained independently and in parallel of each other but are *bagged* (grouped) together into a decision forest to operate as an ensemble.

First, what is a decision tree? As implied by the supervised nature of random forests, decision trees are a type of supervised learners. Decision trees mathematically amount to acyclic graphs that organize nodes into a branching structure. A decision tree is pictured in **Figure 2-6** below.

Each decision tree in **Figure 2-6** is built by recursively partitioning the feature space into non-overlapping segments. The predominant algorithms used to build these trees include ID3, C4.5, C5.0, and CART [62]. The implementation used in this thesis is the CART algorithm, or the *Classification and Regression Trees* algorithm [62, 63]. Per an

implementation's documentation [63], it is possible to break down this decision tree algorithm (CART) into a mathematically succinct formulation. Let:

- X represent the set of feature vectors (in this case, a matrix of k follow-up visits × j features) and let $\boldsymbol{y}$ represent the associated vector of labels.
- the sample set at node m be represented by Q_m and have sample size n_m.
- $\theta_m = \{(j, t) \mid j \in J_m, t \in T_m(j)\}$ represents the set of candidate splits for a node m, where each possible split is defined by one of the features available at the node (J_m), and a specific split point t, where t comes from the set of possible split points for that feature at node m ($T_m(j)$).[1]
- H represent a loss function.

Decision tree building begins at the root node and recursively builds branches by splitting each node's samples until one of the stopping criteria is reached (see below). Thus, beginning at a node m, for each candidate split θ in θ_m, let Q_m be partitioned as

$$Q_m^{left}(\theta) = \{(x, y) \mid x_j \leq t\}$$

and

$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta)\,,$$

where $Q_m^{left}(\theta)$ will contain all of the feature vector - label pairs, (x, y), where feature j's value is within the treshold t. It follows that $Q_m^{right}(\theta)$ will be the subset of data instances from Q_m that are not in $Q_m^{left}(\theta)$. In the CART implementation used in this thesis, the

[1] It is important to clarify that in the CART implementation for the original decision tree, every feature is considered in the set of candidate node splits, and thus there is no explicit distinction for the sets J_m and T_m. In the random forest algorithm, where a random subset of features is considered at each node, this parameterization becomes relevant again, as both the set of features available, J_m, and the set of split points, T_m, will be specific to that node.

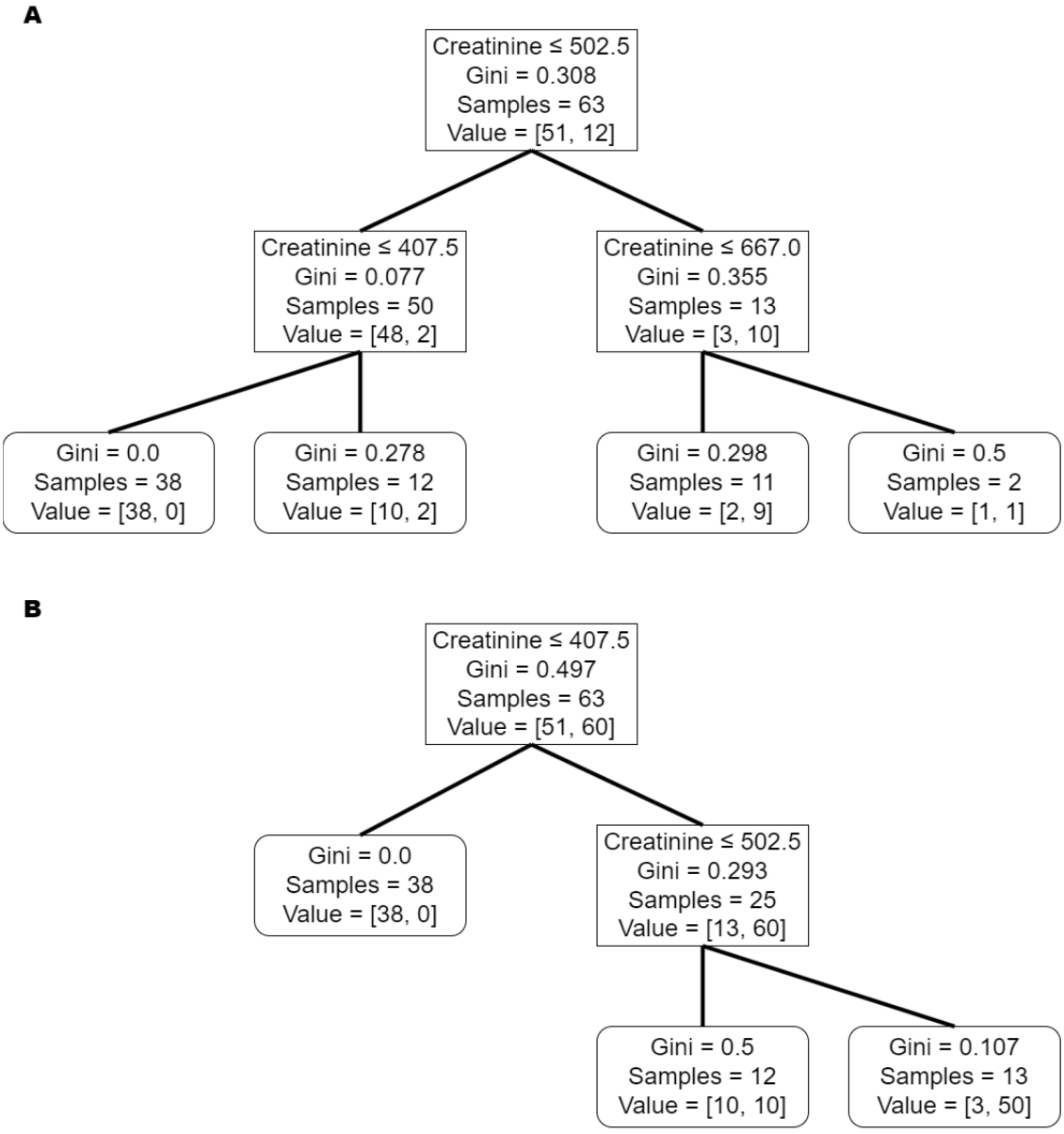


Figure 2-6: Illustration of a decision tree classifier obtained on the patient sample from sections prior. Tree (**A**) is unweighted, while Tree (**B**) had the positive class weighted 5× greater than the negative. Splits are shown in rectangles, while terminal nodes have rounded corners. Creatinine is used as the single feature to partition the data. The *Gini* (impurity metric) of the *Samples* present in each box is given. The class distribution for those samples is given by *Value*, and in the case of Tree (**B**), is 5× weighted for the positive (minority) class.

midpoints between consecutive measurements (sorted) are utilized for each candidate split. For example, if the optimal partitioning at a node occurred between the adjacent

samples of creatinine equal to 404 *μmol/L* and creatinine equal to 411 *μmol/L*, the resulting split point is the midpoint between these samples: 407.5 *μmol/L* (**Figure 2-6B**).

Out of all of the candidate partition pairs yielded from the split set θ_m for the node m, the best split is the one that minimizes the impurity metric H. In this thesis, the impurity metric, or loss function that is used is the Gini information criterion, or the Gini index. It follows that, for node m, the selected parameters θ^* (split feature and threshold), are those that minimize the sum of the impurity criterion in each candidate node's sample set. Concretely,

$$L(Q_m, \theta) = \frac{n_m^{left}}{n_m} H\left(Q_m^{left}(\theta)\right) + \frac{n_m^{right}}{n_m} H\left(Q_m^{right}(\theta)\right) ,$$

where $L(Q_m, \theta)$ is the computed loss (impurity) for the given candidate split θ. And so, the optimal parameters are given by

$$\theta^* = \arg\min_{\theta} L(Q_m, \theta) .$$

The CART algorithm recurses unto θ^* until one of the specified hyperparameters' limits is activated or node m becomes a terminal node with n_m = 1.

As mentioned previously, the criterion selected to be H is the Gini index. For a node m, the Gini impurity is defined as

$$H_{Gini}(Q_m) = \sum_k p_m(k) - p_m(k)^2 ,$$

where

$$p_m(k) = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k) ,$$

and $I(y = k)$ is an indicator function for instances of the k-th class. And so, $p_m(k)$ is the k-th class's proportion in Q_m. An understanding of the above expressions may be facilitated from **Figure 2-7**.

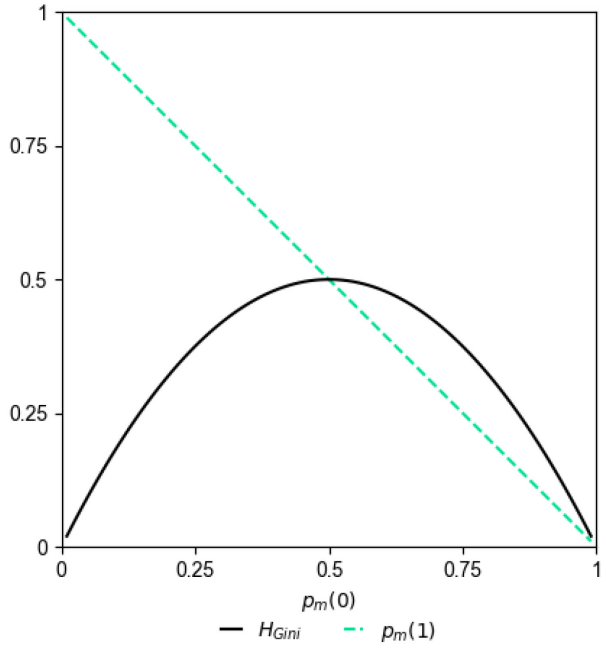


Figure 2-7: Illustration of the Gini metric's range as a function of the positive class proportion, $\boldsymbol{p_m(1)}$, and the negative class proportion, $\boldsymbol{p_m(0)}$, for a hypothetical sample, $\boldsymbol{Q_m}$.

The range and effect of the Gini function is also evident from the decision trees in **Figure 2-6**. Each of the trees in **Figure 2-6A** and **Figure 2-6B** are decision trees fit to the same patient sample from before. Both trees are specified to a maximum depth of two and are fit using a single feature to ensure a consistent result [64]. The difference between Tree A and Tree B, is that Tree B is fit with a weighting applied to instances of the positive class. The positive class is weighted to be 5× the value of the negative class. From this, it may be observed at the root node that the 12 positive instances in the unweighted tree turn into an equivalent weighting of 60 positive instances in the weighted tree. As a result, the

impurity of the node increases, and the resulting tree structure is affected. The Gini values in **Figure 2-6** may be easily calculated by hand using the equations above.

The effect can be further studied by examining the resulting classification metrics for each tree. Respective confusion tables are provided in **Table 2-3** below. It may be illustrated from this example that the weighting improved positive class sensitivity to 83%, compared to the unweighted tree's 75%. Specificity was slightly improved – from 94% to 96%. However, this increased sensitivity to the positive class came at a substantial reduction to precision – from 82% to 77%.

Table 2-3: Confusion tables for the decision trees in **Figure 2-6**, obtained from the data used to fit the trees.

Confusion Table (Tree A)

		Predicted	
		False	**True**
Actual	**False**	49 TN	2 FP
	True	3 FN	9 TP

Confusion Table (Tree B)

		Predicted	
		False	**True**
Actual	**False**	48 TN	3 FP
	True	2 FN	10 TP

Abbreviations: TN, true negative; FP, false positive; FN, false negative; TP, true positive.

At their core, random forests employ decision trees. They *bag* an ensemble of decision trees together to mitigate the instability and overfitting that singular decision trees are prone to. The number of trees to train in each ensemble is a hyperparameter in the algorithm. The fundamental idea that was implemented in random forests to achieve improved performance was to ensure that the individual decision trees were decorrelated from each other. The first measure by which this may be achieved is to train each individual decision tree on a unique sample of the training data. In this case, bootstrap resampling is used. Secondly, as previously mentioned, decision trees in a random forest consider a unique subset of features to split on at each node undergoing a potential split.

This differs to vanilla decision trees where all features are considered for a split. The specific random forest types employed in this thesis are now discussed.

2.3.2.1 Random Forest Classifier

Most of the discussion to this point has been in relation to the random forest classifier. Each tree in the random forest classifier behaves similar to the singular CART decision tree, but with the added modifications of the random selection of features and bootstrapping, discussed above. The output of a trained random forest classifier is a predicted probability of membership to the positive class, in the binary case. This output is an aggregation of the predicted output of each of the singular decision trees in the random forest model.

2.3.2.2 Random Survival Forest

The decision trees, or *survival trees* in a random survival forest behave similar to the random forest + CART algorithms upon fitting [65]. However, within each survival tree, the loss at a splitting node is instead calculated using the log-rank test of the Kaplan-Meier estimate of survival in the two candidate child nodes. The candidate nodes are split to maximize survival dissimilarity between each group, $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$. Predictions are generated in the same manner as the other discussed random forest models. Random survival forest output amounts to the Kaplan-Meier estimate of survival of the individuals (Q_m) in the terminal node of each survival tree, aggregated over each survival tree.

2.4 Conclusion

In **Chapter 4** we evaluate the application of Cox prediction models and machine learning models to predict kidney failure at short timeframes and reduce the incidence of unplanned

dialysis. The objective is to classify patients at each follow-up visit so as to precisely inform the clinician and patient on the imminent risk of kidney failure. The hypothesis is that the specific contexts of this clinical problem demand more specific models than longer-timeframe kidney failure risk prediction models such as the KFRE. Further, the hypothesis is that machine learning models able to account for higher-dimensional data may provide superior short timeframe kidney failure risk prediction. With these models and other data, the clinician and patient can have a timely and informed discussion towards planning a course of action that optimizes survival, quality of life and health care costs while respecting patient wishes. But, before that, in **Chapter 3** the dataset used to establish these models is described in detail.

Chapter 3: Dataset

3.1 Overview

This section provides an overview of the background and preparation of the main dataset used in this study. It highlights where the data was collected, the time period of data collection, and the method of collection. Additionally, this section outlines the specific variables collected in the dataset. It details the types of variables, such as demographic, clinical, and supplementary variables. Finally, it describes data cleaning, imputation, and complementing through feature engineering.

3.2 The Ottawa Hospital Multi-Care Kidney Clinic Dataset

The Ottawa Hospital Multi-Care Kidney Clinic Dataset was collected from the Multi-Care Kidney Clinic (MCKC) at The Ottawa Hospital in Ottawa, Ontario, Canada. The MCKC is a specialized facility that focuses on the treatment and management of kidney diseases and provides advanced care to patients with various kidney conditions. The Ottawa Hospital is a 1,150-bed academic tertiary care center with a catchment area of approximately 1.3 million people and the MCKC is the sole such program within the catchment area. Further information on advanced CKD clinics and the services they provide can be gathered from reading **Section 2.1.5** and **Section 2.1.6**.

The data collection process at Ottawa MCKC was conducted between 2010 and 2021 using a combination of methods, including electronic medical records (EMRs), patient surveys, and nurse and clinician documentation. The EMRs contain detailed information about patients' medical history, diagnoses, medications, laboratory results, and social factors such as education level and marital status. This wide range of variables are meant to capture important patient characteristics and clinical trends to facilitate

optimal monitoring and treatment across different CKD progression types and to empower research in this population [4, 5, 55, 66]. All together, these variables assemble into a detailed and granular dataset encompassing a wide range of data features enabling detailed retrospective analysis of this patient population. This dataset underpins all of the models derived in this thesis.

3.2.1 Contents

The specific contents of the dataset total some 350 unique columns. The most important clinical patient variables are tabulated in **Table 3-1**. Not included in the table are the laboratory measurements previously listed in **Table 2-1** (but are contained in the dataset). Not all variables were used in this work, lending the database for future research.

3.2.2 Ethics and Privacy Compliance

It is important to note that the dataset was de-identified to ensure patient privacy and comply with ethical guidelines. Any personal identifying information, such as names, addresses, specific hospital identification numbers, and precise birthdates were removed or obfuscated to protect patient confidentiality. Throughout the data analysis process, ethical considerations were paramount. The study adhered to strict ethical guidelines to ensure the protection of patient privacy, confidentiality, and data security. Additionally, the study obtained appropriate approvals from relevant institutional review boards and complied with all legal and regulatory requirements. Specifically, all protocols were approved by the Ottawa Health Science Network Research Ethics Board (Protocol ID #20150457-01H). Informed consent was waived due to the retrospective nature of the dataset, and strict data access controls were implemented to prevent unauthorized use or disclosure of sensitive information. Any potential conflicts of interest were disclosed and

managed appropriately to maintain the integrity and objectivity of the findings. Rigorous data validation and verification procedures were employed to ensure the accuracy and reliability of the results.

Table 3-1: Key patient variables used in subsequent analyses (in addition to those in **Table 2-1**).

Outcome Variables	**Details**
Date of death	Rounded to first day of the month
Date of kidney replacement therapy	Rounded to first day of the month
Kidney replacement therapy modality	Peritoneal / hemodialysis
Inpatient / outpatient dialysis	Urgent dialysis requiring hospitalization, home dialysis, etc.
Demographics	
Age	From date of birth and time of visit
Sex	Male/Female
Race	Self-defined
Disease Etiology and Comorbidities	
CKD type	
Diabetes	Yes/No
Hypertension	Yes/No
Coronary artery disease	Yes/No
Congestive heart failure	Yes/No
Anthropometrics and Vital Signs	
Heart rate	At rest
Systolic and diastolic blood pressure	At rest
Body mass index	
Medications	
Diuretics	Yes/No and dates prescribed
ACE inhibitors and ARB	Yes/No and dates prescribed
Other Important Variables	
Dates of visits with clinic MD	Used to anchor laboratory values
Opted for conservative care management	An exclusion criterion

Abbreviations: ACE, angiotensin-converting-enzyme; ARB, angiotensin II receptor blocker; MD, clinician / nephrologist.

3.2.3 Preprocessing

Thorough preprocessing was paramount to ensure data quality and consistency. This phase occurred early in the project and involved a succession of steps including data cleaning and handling of missing data.

- *Data cleaning*:
    - Aimed to identify and rectify manual-entry errors, inconsistencies, or outliers present in the dataset. It involved scrutinizing the variables for improbable values, data entry mistakes, and inconsistencies in coding. When errors or outliers were detected, appropriate actions, such as correction from original data sources, imputation, or removal, were taken to rectify them. The dataset underwent two rounds of manual review with administrative clinic nurses.
- *Missing data:*
    - Depending on the extent and nature of the missingness, various techniques were employed to maximize data availability. The specifics of the imputations performed are discussed in **Section 3.4**. Under extenuating circumstances, entire patient series or variables were excluded.

3.2.4 Limitations

Several important limitations may be identified with respect to the dataset. Specifically discussed are the particularities of this dataset, and their potential implications.

Firstly, the generalizability of this dataset and the resultant models has yet to be tested in diverse healthcare settings. This dataset characterizes a cohort obtained from a single healthcare center, namely The Ottawa Hospital's MCKC. The findings presented herein may therefore not be directly generalizable to other healthcare settings or patient

populations, but regardless are presumed to be representative of current local practice. Nevertheless, **Chapter 4** and **Chapter 5** demonstrated that the derived models validate well in external healthcare sites within Ontario where practice is similar. Practices further abroad – in other provinces or nations – may differ more significantly. The generalizability of findings may be jeopardized in these global settings. It is important to consider the specific context and characteristics of the dataset when interpreting the results.

Second, missing data is a particular nuisance in this dataset due to the unfortunate collection timeline. Urine ACR has been continually validated to be a prime predictor of kidney function, warranting its inclusion in most kidney failure prediction models [40], and leading to a mandate for its collection in advanced CKD clinics by the Ontario Renal Network (ORN). This mandate came into effect between 2015 – 2016, and the effect of this will become apparent in **Section 3.3**. As such, the current dataset contains an artifact (high and non-random ACR missingness) that is difficult to mitigate. Had data collection commenced on 2016 or later, this would not be an issue. This characteristic, and more generally data missingness, place important constraints on the interpretations contained within this thesis.

On this note, as with any observational study, the presence of unmeasured or unaccounted confounding factors may influence the observed relationships between variables. Once again, urine ACR is at the center of the issue. As mentioned in sections prior, medications are often administered to patients in order to manage hypertension and lower proteinuria (elevated levels of protein in the urine). This results in artificial fluctuations in a patient's measurement series as demonstrated in the patient series in **Figure 3-1**. The figure shows the series of urine ACR measurements for the patient from our selected sample and the period over which ARB medication was prescribed. This is likely to be directly responsible for lowering the patient's urine ACR values. It is difficult to

correct this confounding with simple binary indicators and is a direction for future work (**Section 6.3**). In summary, confounding is present in this dataset and should be acknowledged.

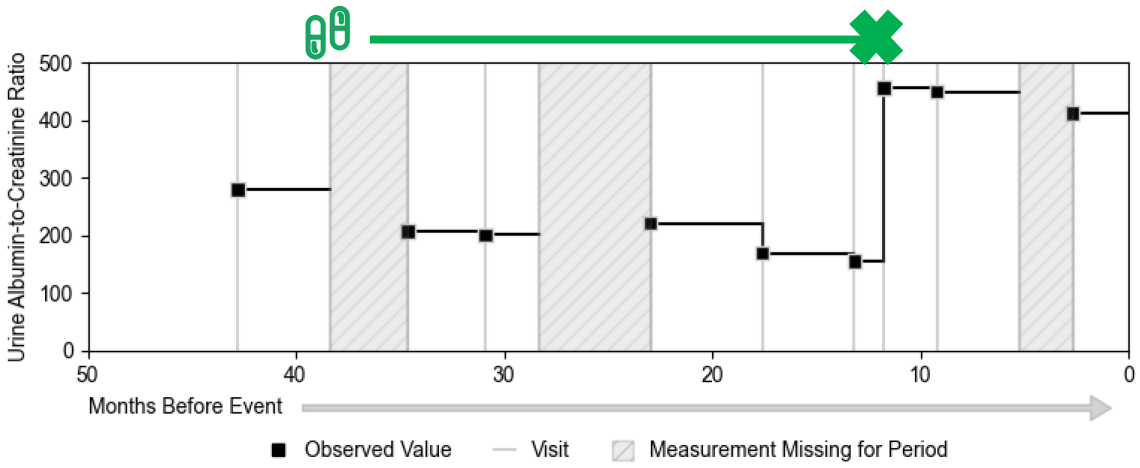


Figure 3-1: Patient urine-albumin-to-creatinine ratio measurements and duration of ARB prescription. The patient's time on the ARB medication is annotated in green.

While extensive data cleaning and validation were performed, the dataset's quality is contingent upon the accuracy and completeness of the original data sources. Errors or inconsistencies in documentation or data entry may still be present, which could impact the validity of any analyses.

Another potential limitation is the inclusion of data from multiple regional laboratories, which may use different methods for measurement. This source of variability may be interpreted as a source of biases in certain patients. Alternatively, this may be seen as capturing a wider range of data that may improve generalization to other practices.

Finally, the time period over which the data were collected may have implications for any analysis, as medical practices, treatment protocols, and patient outcomes may evolve in the future. Once again, the findings should be interpreted within the context of the specific time period in which the data were collected.

3.2.5 Other Datasets Used in This Thesis

Two additional datasets were obtained from independent patient cohorts. One cohort was obtained from Toronto's Sunnybrook Hospital – part of the University Health Network (UHN). Another cohort was obtained from the Kingston General Hospital's (KGH) advanced CKD patient group. Both patient populations are defined as advanced CKD cohorts. Given their location in Ontario, practices across these centers can be considered relatively uniform.

A tabulation of important clinical characteristics for all three cohorts is in **Table 3-2** of **Section 3.3.4**. In summary, patients in the KGH and UHN cohorts were on average in the earlier stages of their CKD progression. In the table, this manifests as increased eGFR, decreased creatinine, reduced proteinuria, lower outcome rates, etc. This has the potential to distort the performance results due to lower outcome/label prevalence. It does not, however, inherently represent a barrier to successful application of a TOH-derived model to these external cohorts.

3.3 Characteristics

This data characteristics section delves into various aspects that describe the dataset and cohort. It highlights the time of entry for patients into the cohort and the duration of follow-up, providing insights into the longitudinal nature of the dataset. This section also explores the nature of patient follow-up and any changes or trends in cohort characteristics over time. Furthermore, it discusses the presence of missing data and its impact on dataset integrity.

3.3.1 Patient Time of Entry and Duration of Follow-up

As previously mentioned, patients entered the clinic and were monitored asynchronously over the specified time period (2010 – 2021). The dataset generally includes all such patients who were referred to the clinic during this time. Timing of referral is at the discretion of the primary nephrologist, though referrals are suggested when the estimated glomerular filtration rate (eGFR) is <25 *mL/min/1.73m²* or the 2-year 4-variable KFRE score is >20% [10, 49]. Patients are typically seen in the clinic every three months, though this interval can vary from as often as every two weeks to as long as every six months per the discretion of the nephrologist. The median and mean survival times are approximately 19 and 25 months respectively. These quantities represent the median and mean duration of follow-up for all of the observed patients in the dataset. The median survival time, stratified among several groups, can be gleaned from **Figure 3-2 A-B** by finding each curve's point of intersection with $\hat{S}(t) = 0.5$.

The duration of follow-up for each patient is defined as the period between the time of first referral and the time of outcome event, or the time of data collection for those patients still being followed. The time of entry represents the date when each patient's data was initially recorded within the cohort. Follow-up durations among patients varied due to a variety of factors. For example, the median duration of follow-up for patients presenting to the clinic with an initial creatinine over 300 *μmol/L* was than half that of patients presenting to the clinic with an initial creatinine on the other side of that threshold (**Figure 3-2A**). Other factors may influence patient survival time. For example, patients opting for conservative care management choose not to receive kidney replacement therapy in the form of dialysis or kidney transplantation. However, they continue to be followed in the clinic so that their quality of life and health can be maintained as much as

possible. These patients exist in large proportion in this dataset (roughly 400 patients). They are almost always excluded from analyses in kidney failure prediction.

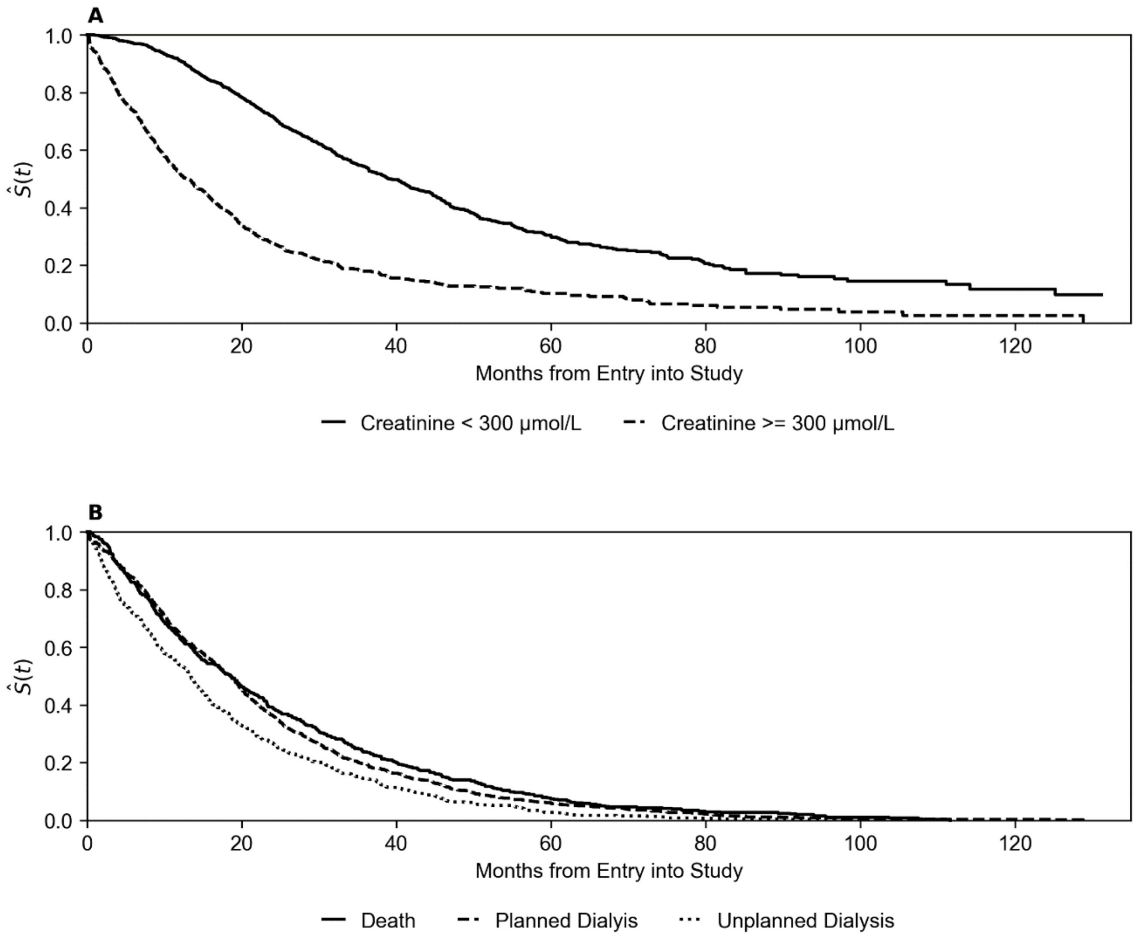


Figure 3-2: Kaplan-Meier estimate of survival among different patient groups.

3.3.2 The Nature of Patient Follow-up

The nature of patient follow-up provides valuable insights into the frequency and regularity of data collection. In this dataset, patient follow-up visits were scheduled at regular intervals, typically every 3 – 6 months. During each visit, various clinical measurements and assessments were recorded. Which variables were collected, and the practice surrounding their collection has been discussed in prior sections. Ultimately, the dataset

is a collection of short, irregularly sampled time series with a multitude of clinical predictors anchored to each time point. For the TOH MCKC, there were, on average, 6-7 follow-up visits (data time points) for each patient in the dataset. This number can vary: many patients with a single visit exist in the dataset, as do patients with over 20 visits. Of the patients contained within the dataset, most experienced an outcome event (death, dialysis, or kidney transplantation). But a subset of patients was still being followed at the time of data collection.

3.3.3 Outcomes

As mentioned in **Section 2.1.6**, outcomes can be broadly classified into either kidney failure, or death. Here, kidney failure is defined as kidney replacement therapy (KRT), meaning the initiation of dialysis or kidney transplantation. Within this category, two subgroups are differentiated for the purposes of this thesis. They are, dialysis initiation in an unplanned setting (unplanned dialysis; UD), and dialysis initiation in the planned setting (planned dialysis; PD). As previously mentioned, unplanned dialysis is best defined as initiating dialysis in the inpatient setting (i.e., the patient presented to the emergency room and was admitted for hospital care). Planned dialysis is then defined as initiating dialysis in the outpatient setting or pre-emptively undergoing kidney transplantation. The incidences for these outcomes, as well as the outcome of death, are plotted with respect to time in **Figure 3-3**.

In the context of predicting kidney failure, death may be treated in different ways. In certain cases, death may have resulted from kidney failure, and therefore could be considered as a surrogate for unplanned dialysis. However, death may also result from an accident, in which case it may be considered as a censoring event. The cause of death, however, is not always so obvious, as patients may have comorbidities or the cause of

death is not documented. In some contexts, death may be treated as a competing event (with kidney failure defined as kidney replacement therapy) [47, 67, 68]. In the context of this work, death is treated in accordance with the data labeling policy that was employed. I.e., “is the patient within 6 or 12 months of a kidney failure event defined as the initiation of kidney replacement therapy?” The drawbacks of this approach are discussed in **Chapter 6** as an avenue for further research.

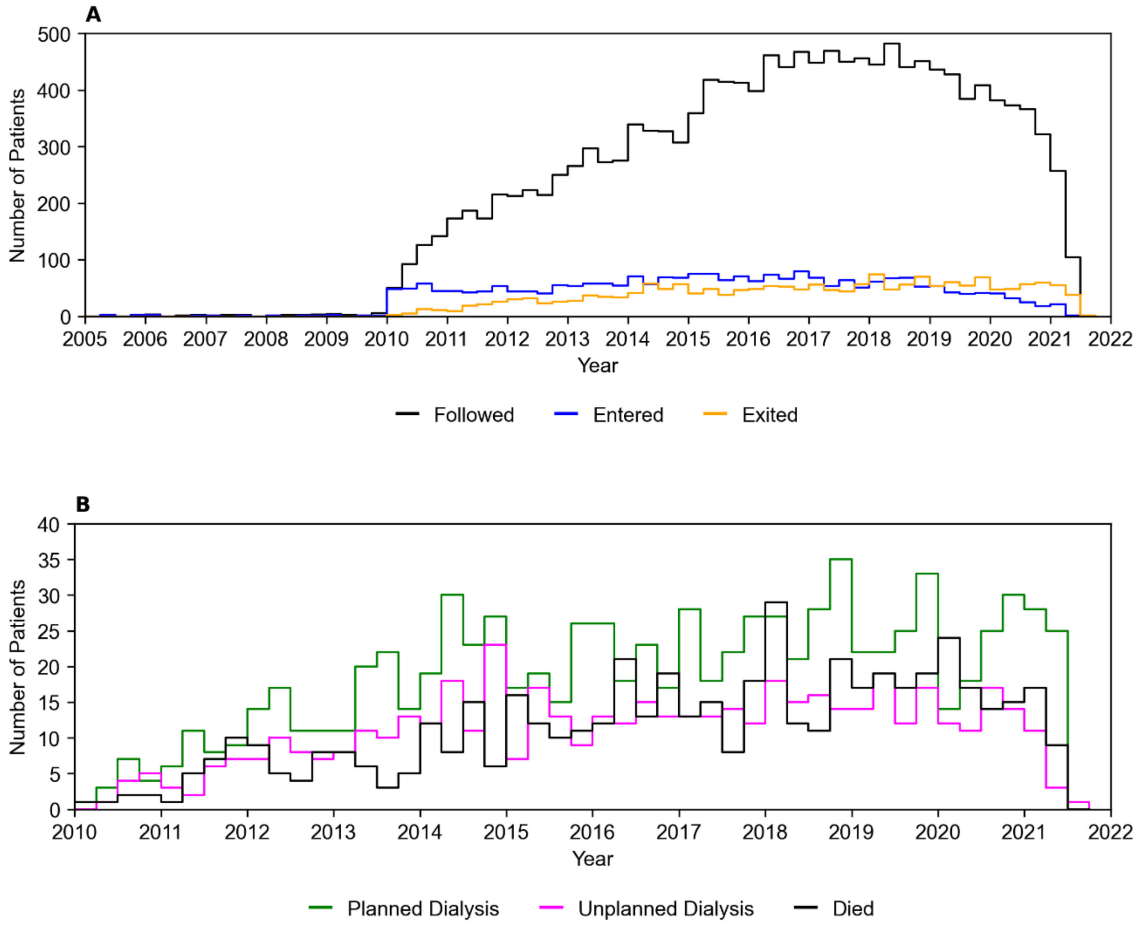


Figure 3-3: The Ottawa Hospital Multi-Care Kidney Clinic cohort patient numbers over time. Panel (**B**) is the breakdown data of the *Exited* group in panel (**A**). Sampling is performed quarterly.

3.3.4 Variables

Variable distributions are available in **Table 3-2**. The characteristic worth highlighting with respect to the dataset variables is the high missingness in a number of important clinical characteristics (**Figure 3-4**). The most important missing variable is uACR, which as previously mentioned, only started being required to be collected in 2015. In 2020 and 2021 a higher rate of missing blood pressure (BP) measurements was noted, associated with greater prevalence of virtual visits in response to the COVID-19 pandemic. The impact of this data characteristic could be profound in either data loss or added bias. For example, the magnitude and nature of the missingness in measured diastolic and systolic blood pressure in the KGH external cohort necessitated the exclusion of these variables from the analyses. This was because this covariate data was missing in 50% of cases, but more importantly, was only filled in in the later times of each patient series. An analysis of missing variables in the TOH cohort is available in **Section 3.4**.

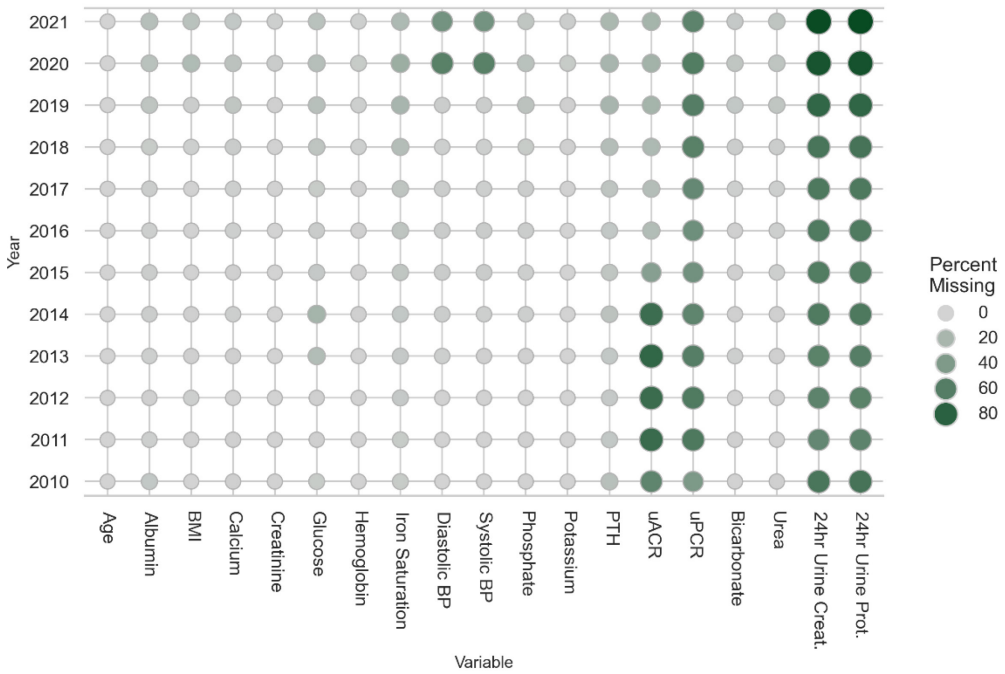


Figure 3-4: Variable missingness across the available patient data, with year of data collection.

Table 3-2: Baseline characteristics of study cohorts.

Characteristics	TOH (N = 1849)	%	KGH (N = 1033)	%	UHN (N = 323)	%
Demographics						
Age, Mean (SD)	66 (15)	0	* 70 (14)	0	68 (17)	0
Female Sex, N (%)	690 (37)	0	401 (39)	0	129 (40)	0
Laboratory Data						
Creatinine, Mean (SD)	308 (97)	0.9	* 277 (97)	3.5	* 252 (87)	0
eGFR, Mean (SD)	19 (7)	0.9	* 21 (7)	3.5	* 23 (8)	0
uACR, Median (IQR)	162 (49, 333)	32.2	* 109 (29, 282)	13.4	* 84 (27, 219)	0.2
Calcium, Mean (SD)	2.23 (0.15)	3.7	* 2.29 (0.18)	7.7	* 2.30 (0.14)	1.3
Phosphate, Mean (SD)	1.37 (0.31)	4.4	* 1.32 (0.31)	7.9	* 1.30 (0.27)	1.5
Bicarbonate, Mean (SD)	24 (3)	3.5	* 23 (4)	7.1	* 23 (3)	1.2
Potassium, Mean (SD)	4.5 (0.6)	1.2	4.6 (0.6)	3.8	* 4.7 (0.6)	0.9
Albumin, Mean (SD)	35 (5)	4.7	35 (6)	8.5	* 40 (5)	1.6
Comorbidities						
Diabetes, N (%)	1110 (60)	0	634 (61)	0	185 (57)	0
Hypertension, N (%)	1689 (91)	0	909 (88)	0	307 (95)	0
CHF, N (%)	408 (22)	0	* 170 (16)	0	* 40 (12)	0
Vital Signs / Anthropometrics						
Systolic BP, Mean (SD)	137 (21)	9.1	136 (21)	36	134 (21)	37.5
Diastolic BP, Mean (SD)	72 (13)	9.4	73 (12)	36	74 (13)	37.5
BMI, Mean (SD)	29.9 (7.1)	3.3	* 31.3 (7.9)	51.8	29.5 (12.0)	16
Outcomes						
Being followed, N (%)	281 (15)		363 (35)		182 (56)	
Kidney transplantation, N (%)	91 (5)		10 (1)		8 (2)	
Planned (outpatient) dialysis, N (%)	682 (37)		300 (29)		63 (20)	
Unplanned (inpatient) dialysis, N (%)	435 (24)		161 (16)		22 (7)	
Died in predialysis, N (%)	360 (19)		199 (19)		48 (15)	

Abbreviations: SD, standard deviation; IQR, inter-quartile range; eGFR, estimated glomerular filtration rate; N, number; TOH, The Ottawa Hospital; KGH, Kingston General Hospital; UHN, University Health Network Toronto; %, percent missing from dataset; uACR, urine albumin-to-creatinine ratio; eGFR, estimated glomerular filtration rate; CHR, congestive heart failure; BMI, body mass index; BP, blood pressure.
*: Significantly different from TOH at p<0.01. Welch's Test is used for continuous covariates where the assumptions of normality are met. Mann-Whitney U Rank Test is used for continuous covariates that are not normally distributed. A X^2 test is performed for binary covariates.

3.4 Missing Data

The high missingness in key predictors, namely urine albumin-to-creatinine ratio (uACR) (shown in **Section 3.3**), introduces additional complexity and risk for bias into a model. Laboratory measurements such as uACR are frequently missing and not at random [69], as will be further evidenced in the coming sections. This makes the complete mitigation of bias difficult, and likely impossible. Nevertheless, being a significant predictor, ignoring of uACR is detrimental. To this end, two main approaches are immediately evident. The first involves excluding those patients entirely. But, this was ruled out for two reasons: (1) it would reduce the patient cohort to nearly half, thereby increasing the risk for model overfitting, and (2) doing so would bias the sample population in terms of sex, bicarbonate, calcium, phosphate (data shown in **Table 3-6**).

An alternative approach is to impute missing values based on other corelates within the dataset. In this section, several imputation methods are presented and evaluated for two types of scenarios: 1) interpolating inner measurement series values (those missing values in-between one or more observed measurements), and 2) for baseline value imputation (missing values occurring at the beginning of a patient measurement series).

Save for one of the imputation methods to be presented, the approaches considered here operate within the individual patient series. Time-series data often exhibit autocorrelation, meaning that earlier values can explain or contribute information on later values. This is especially true for many of the covariates present in this clinical dataset. Additionally, in the context of this clinical dataset where variables are frequently confounded by outside factors such as medications, imputing using information local to the patient makes intuitive sense for accurate prediction of missing values over population, or imputation using models of relationships with other patient-specific covariates from the

same timepoint. This section discusses the imputation methods applied to the datasets utilized in these analyses, including those of **Chapter 4**.

3.4.1 Imputation Methods

Two primary types of missing values may be identified in these data. Missing values may be neighbored by either one or two observed measurements. For example, in **Figure 3-5B**, the patient is missing a uACR measurement at initial visit ($t = 43$ months before event) and is thus neighbored by only one observed data point at $t = 35$ months. Continuing through the series, more missing values pop up intermittently between recorded observations. Imputing data between neighboring points allows the imputation method to leverage multiple data points (interpolate), potentially reducing imputation error. Projecting outwards (extrapolating) involves making assumptions about the data's trend. It follows that these methods are generally more volatile and error prone. **Figure 3-6** illustrates the two approaches studied for imputing missing values neighbored by two observed measurements. They are discussed in the coming sections. Note that the measurement timelines in **Figure 3-5** to **Figure 3-7** are illustrated in a piecewise-constant manner to be in line with the manner time-varying Cox models typically treat the space in between measurements.

3.4.1.1 Last Observation Carried Forward (LOCF)

Last observation carried forward (LOCF) is a naïve approach to data imputation that is generally reported to produce misleading and erroneous values [70]. The procedure is simple: carry forward the most recently observed measurement value (**Figure 3-6A**). It is worth noting that data characteristics vary, and thus so too will the optimal imputation method. In the case of a dataset composed of many short and noisy time series

measurements, a simplistic approach has the added benefit of interpretability and low variance. It is also worth noting that although this approach is used to fill in data missing in between neighboring measurements, it is fundamentally an extrapolation process from the last observation.

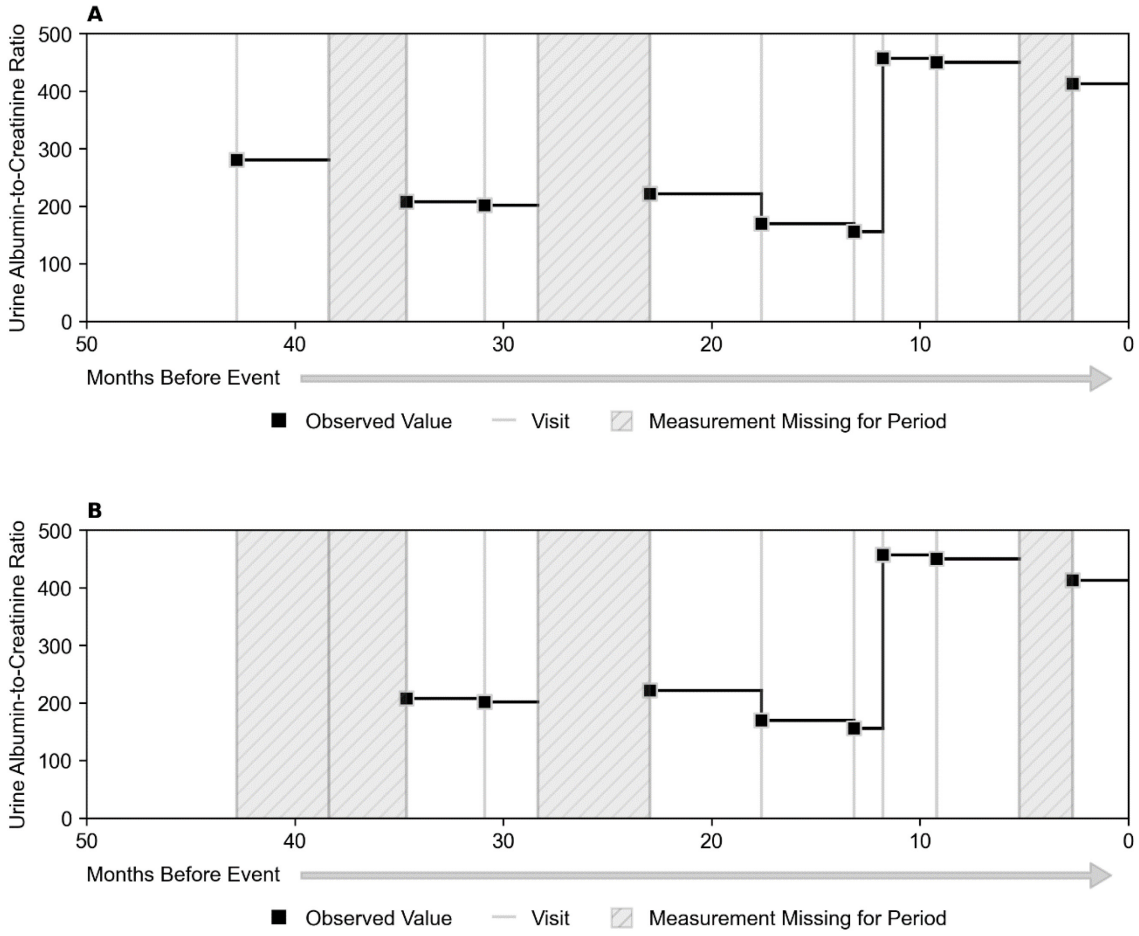


Figure 3-5: Illustration of a series of urine albumin-to-creatinine ratio (ACR) measurements for a single patient. In (**A**), the original series is shown. In (**B**), the baseline measurement is intentionally dropped for the experiment where several baseline imputation methods are compared (**Section 3.4.2**).

3.4.1.2 Time-Scaled Linear Interpolation

Time-scaled linear interpolation (**Figure 3-6B**) involves connecting neighboring measurements with a line and imputing missing values at the appropriate time points along

the line. It is equivalent to a time-weighted average between neighboring points. This method therefore uses two data points to perform imputation, as opposed to the single datum used in LOCF, and assumes a linear trend between the two adjacent measurements.

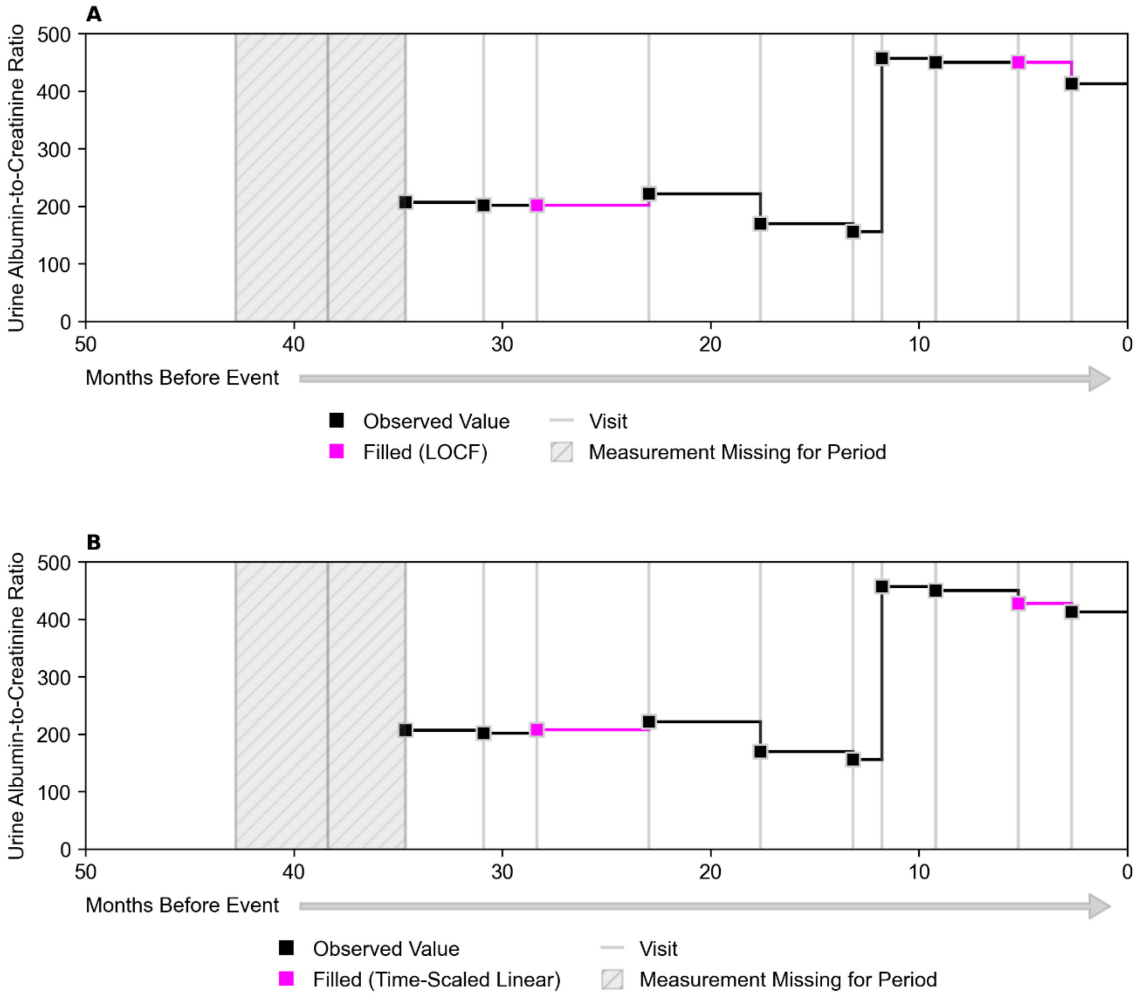


Figure 3-6: Example of the two interpolation strategies being performed on urine albumin-to-creatinine ratio (uACR) measurements for one of the patients from the selected patient sample from sections prior. Panel (**A**) demonstrates the *last observation carried forward* (LOCF) approach. Panel (**B**) demonstrates a time-scaled linear interpolation approach. Note that the difference between the imputed values in panels (**A**) and (**B**) is slight in this particular patient.

3.4.1.3 Next Observation Carried Backward (NOCB)

The first method for baseline imputation is the naïve method of *carrying the next observation backward* (NOCB). Simply put, the closest future observation is carried backward to impute the missing baseline value (**Figure 3-7A**). It is usually met with many of the same criticisms as LOCF.

3.4.1.4 Sex-Stratified Median

Another naïve imputation method involves computing the median of the specific laboratory value being imputed, for both sexes, and imputing the baseline measurement that way. This method, while interpretable, introduces values likely to be decorrelated from the patient's measurement series. It is illustrated in **Figure 3-7B**.

3.4.1.5 Multiple-Fixed Linear Regression

This imputation method performs N local linear regression imputations within an individual's measurement series, where N is the number of observed data points. Each of the local regressions is fixed to one of the observed data points and fit to the remaining data points. The imputed baseline value is the aggregation (mean) of the imputations obtained from each of the N local regressions. The result of this method is illustrated in **Figure 3-7C**.

3.4.2 Analysis

An analysis was performed to quantify the potential biases that would be introduced by each imputation method. Two simulations were performed. Simulation one (interpolation) involved randomly dropping from the inner parts of the measurement series. Tests were performed for both a single random drop, and a double random drop. Simulation two

(extrapolation) involved dropping the baseline measurement of each patient that had an observed baseline measurement and at least two other observed data points. Imputation was then performed according to the methodology, and predicted values were compared against the true observed values using metrics for the mean of the absolute errors (MAE), and the root of the mean squared errors (RMSE). Confidence intervals are obtained by taking the 2.5 and 97.5 percentiles (95% CI) of 1,000 bootstrap resamples of the results.

Table 3-3 and **Table 3-4** tabulate the performance results for each imputation method across a suite of selected laboratory measurements, where either a single drop or a double drop was performed, respectively. **Table 3-5** tabulates the analysis results across baseline imputation methods for the same laboratory measurements. In each table, cells are colored according to the relative ranking of the performance result within each metric, and within each laboratory measurement. Better results are colored more boldly.

Associations in missing baseline variables were assessed using a test for data missingness at random (MAR). This test sought to characterize the likelihood of observing a missing data point based on the observed values of other variables. The available clinical characteristics were included as independent covariates into a logistic regression model. Supplementary covariates for time of the year were also included. The dependent variable was missingness indicator for the covariate being studied. A second test was performed to assess the independence between the missingness in one covariate to all of the other covariates. A X^2 test was performed on a contingency table comparing missingness in the covariate being studied to missingness among the other covariates. These results are tabulated in **Table 3-6**.

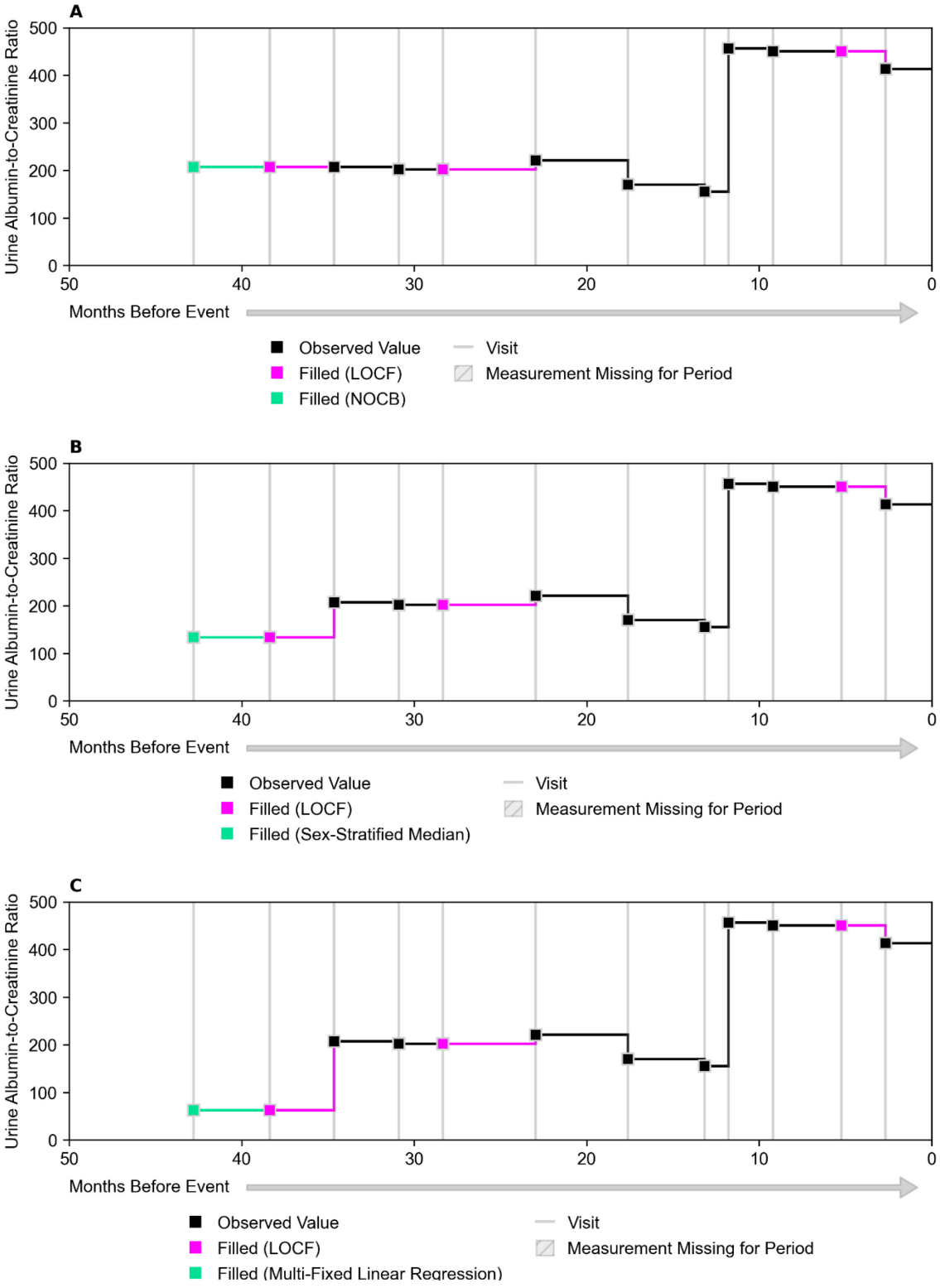


Figure 3-7: Illustration of the three baseline imputation approaches that were considered. Baseline-imputed values are colored emerald. The LOCF-filled measurements are colored fuchsia, as before. Panel (**A**) illustrates the *next observation carried backward* (NOCB) approach. Panels (**B**) and (**C**), depict a sex-stratified median imputation and a multiple-fixed linear regression imputation, respectively.

3.4.3 Discussion

The results of the imputation analysis generally indicate that in most of the studied cases, missing values can be imputed with reasonable accuracy using straightforward imputation approaches. For example, the NOCB MAE of 39 *μmol/L* obtained for baseline imputation on creatinine represents a mean absolute error of 11% over the average advanced CKD patient at their initial visit to the clinic (350 *μmol/L*). This represents just a 4% increase over the natural intra-day variability experienced by patients in this cohort [71].

Overall, the lowest error for the task of extrapolating to baseline (initial visit) was achieved using the naïve NOCB approach (**Table 3-5**). This may be the case for two reasons: (1) the sex-stratified median imputed values are decorrelated from the local patient data, and (2) the time series data are too noisy and too short on average to benefit the multiple linear regression (MFLR) approach. While the NOCB approach statistically significantly outperformed both of the other methods, the difference between NOCB and MFLR is likely not clinically significant. The same may be said for the interpolation results (**Table 3-3** and **Table 3-4**): while time-scaled linear interpolation statistically significantly outperformed LOCF in terms of MAE and RMSE, the difference is likely not clinically significant (e.g., difference in imputed vs. actual MAE for creatinine equals *10 μmol/L* between the two studied methods). Given the marginal improvement of linear interpolation over LOCF, other considerations were taken into account. For example, LOCF is the only of the two approaches that would be applicable to the most recent clinical visit. Therefore, LOCF would more closely mimic the imputation process that would be required at the most recent patient visit in a prospective clinical setting.

In summary, the results of this analysis supported the move to include patients missing these laboratory measurements in the modeling by using NOCB imputation for missing baseline (initial visit) measurements and LOCF for all other imputations. Doing so

allowed for the inclusion of those upwards of 18% of patients missing crucial laboratory measurements such as uACR, and thus a significant portion of the dataset was retained for modeling.

Table 3-3: Interpolation imputation analysis for selected laboratory measurements (single drop).

	MAE (LOCF)	RMSE (LOCF)	MAE (Linear Interp.)	RMSE (Linear Interp.)	N
Creatinine	43 (42, 45)	51 (49, 52)	33 (31, 34)	38 (36, 39)	1615
Urine ACR	74 (69, 79)	87 (81, 92)	57 (53, 61)	66 (62, 70)	1134
Calcium	0.10 (0.09, 0.10)	0.11 (0.11, 0.12)	0.08 (0.08, 0.08)	0.09 (0.09, 0.10)	1582
Phosphate	0.21 (0.21, 0.22)	0.25 (0.24, 0.26)	0.18 (0.18, 0.19)	0.21 (0.20, 0.22)	1581
Albumin	2.93 (2.84, 3.02)	3.49 (3.40, 3.59)	2.43 (2.35, 2.50)	2.80 (2.72, 2.88)	1575
Potassium	0.43 (0.42, 0.44)	0.50 (0.49, 0.51)	0.36 (0.35, 0.37)	0.42 (0.41, 0.43)	1616
Bicarbonate	2.33 (2.26, 2.40)	2.73 (2.65, 2.80)	2.01 (1.94, 2.07)	2.29 (2.22, 2.36)	1588

Abbreviations: MAE, mean of the absolute error; RMSE, root of the mean squared error; urine ACR, urine albumin-to-creatinine ratio; Linear Interp., time-scaled linear interpolation; N, number.
Color: Better results (compared by metric and laboratory measurement) are indicated in dark green cells.

Table 3-4: Interpolation imputation analysis for selected laboratory measurements (double drop).

	MAE (LOCF)	RMSE (LOCF)	MAE (Linear Interp.)	RMSE (Linear Interp.)	N
Creatinine	48 (47, 50)	58 (56, 60)	35 (34, 36)	41 (40, 43)	1424
Urine ACR	77 (72, 83)	93 (87, 99)	60 (56, 64)	71 (67, 76)	952
Calcium	0.10 (0.10, 0.10)	0.12 (0.11, 0.12)	0.08 (0.08, 0.09)	0.10 (0.10, 0.10)	1393
Phosphate	0.22 (0.21, 0.22)	0.26 (0.25, 0.26)	0.19 (0.18, 0.19)	0.22 (0.22, 0.23)	1384
Albumin	3.03 (2.95, 3.12)	3.67 (3.57, 3.77)	2.49 (2.42, 2.56)	2.97 (2.89, 3.05)	1382
Potassium	0.44 (0.43, 0.45)	0.52 (0.50, 0.53)	0.37 (0.36, 0.38)	0.44 (0.43, 0.45)	1424
Bicarbonate	2.41 (2.35, 2.48)	2.88 (2.81, 2.95)	2.06 (2.01, 2.12)	2.43 (2.37, 2.49)	1393

Abbreviations: MAE, mean of the absolute error; RMSE, root of the mean squared error; urine ACR, urine albumin-to-creatinine ratio; LOCF, last observation carried forward; Linear Interp., time-scaled linear interpolation; N, number.
Color: Better results (compared by metric and laboratory measurement) are indicated in dark green cells.

Table 3-5: Baseline imputation analysis for selected laboratory measurements. Results are mean (95% CI).

	MAE (Sex-Stratified Median)	RMSE (Sex-Stratified Median)	MAE (NOCB)	RMSE (NOCB)	MAE (MFLR)	RMSE (MFLR)	N
Creatinine	70 (68, 72)	86 (83, 90)	39 (37, 41)	54 (51, 58)	62 (58, 65)	97 (84, 109)	1611
Urine ACR	160 (150, 172)	243 (223, 262)	89 (82, 96)	150 (134, 164)	129 (119, 140)	218 (200, 236)	1054
Calcium	0.11 (0.10, 0.11)	0.14 (0.14, 0.15)	0.10 (0.09, 0.10)	0.13 (0.12, 0.14)	0.12 (0.11, 0.12)	0.18 (0.16, 0.21)	1567
Phosphate	0.22 (0.21, 0.23)	0.29 (0.27, 0.30)	0.22 (0.21, 0.23)	0.29 (0.28, 0.31)	0.27 (0.25, 0.29)	0.43 (0.38, 0.49)	1562
Albumin	3.99 (3.83, 4.16)	5.20 (4.98, 5.41)	3.58 (3.45, 3.73)	4.62 (4.43, 4.84)	4.24 (4.06, 4.43)	5.65 (5.35, 5.97)	1550
Potassium	0.46 (0.44, 0.48)	0.57 (0.55, 0.59)	0.45 (0.43, 0.46)	0.59 (0.56, 0.62)	0.54 (0.51, 0.57)	0.84 (0.69, 1.08)	1609
Bicarbonate	2.69 (2.58, 2.81)	3.46 (3.31, 3.62)	2.34 (2.24, 2.45)	3.12 (2.98, 3.27)	2.98 (2.78, 3.25)	5.46 (3.83, 7.57)	1562

Abbreviations: MAE, mean of the absolute error; RMSE, root of the mean squared error; urine ACR, urine albumin-to-creatinine ratio; NOCB, next observation carried backward; MFLR, multi-fixed linear regression; N, number; 95% CI, 95% confidence interval.
Color: Better results (compared by metric and laboratory measurement) are indicated in darker green cells.

Finally, several consequential associations in baseline visit variable missingness were uncovered. They are shown in **Table 3-6**. Most of the missingness was present in patients' uACR samples. As the largest source of missingness in this dataset, this potentially represents a major source of bias that ideally will be rectified in downstream analyses and the usage of more recent and complete datasets.

Table 3-6: P-values for the results of variable missingness analysis.

	Creat.	uACR	Calc.	Phos.	Alb.	Potass.	Bicarb.	Female Sex	Age	Year	Quarter	X^2	N
Creatinine	NA	0.051	0.293	0.258	0.502	0.611	0.062	0.095	0.636	**0.002**	0.208	**0**	9
Urine ACR	**0**	NA	**0.024**	**0**	0.243	0.874	**0.028**	**0.001**	0.204	**0**	0.873	**0**	317
Calcium	0.111	0.907	NA	0.280	**0.029**	0.902	0.608	0.414	0.055	**0**	0.195	**0**	48
Phosphate	**0.017**	0.692	0.156	NA	0.556	0.389	0.444	0.605	0.109	**0**	0.121	**0**	48
Albumin	**0.006**	0.395	0.203	0.176	NA	0.449	0.323	0.631	0.431	**0**	0.088	**0**	66
Potassium	0.501	0.376	0.560	0.553	0.198	NA	0.966	0.159	0.534	**0**	0.336	**0**	12
Bicarbonate	0.820	**0.039**	0.226	0.101	0.399	0.497	NA	0.146	0.118	**0**	0.307	**0**	50

Abbreviations: NA, result is not-applicable; uACR, urine albumin-to-creatinine ratio; creat., creatinine; calc., calcium; phos., phosphate; alb., albumin; potass., potassium; bicarb., bicarbonate; X^2, Chi-Square Test; N, number missing. P-values < 0.05 are bolded.
Interpretation: Table is read from left-to-right. For example: missingness in creatinine is not associated with elevated urine albumin-to-creatinine ratio to a significant degree (p-value = 0.051) based on the MAR test that was conducted.

3.5 Feature Engineering

Laboratory measurements play a critical role in the clinical management of many chronic conditions, providing valuable insights into disease progression and response to potential interventions. As previously mentioned, clinicians monitoring patients with advanced CKD prefer to collect multiple laboratory samples at routine and frequent intervals instead of relying on only a single set of measurements taken at a patient's initial visit. This is evidenced by the existence of these repeated measurements in the dataset. In clinical practice, examining the trends in these series, such as the changes in values over time, allows clinicians to make more timely and informed decisions.

Despite the recognized importance of analyzing trends in laboratory measurements in monitoring CKD progression, current kidney failure prediction models have not fully incorporated these data. Most prediction models for CKD progression primarily rely on baseline measures of kidney function, demographic data, and comorbidities [39, 40]. One model, developed by Tangri *et al.*, used repeated laboratory measurements to build a time-varying Cox model [72]. This model can dynamically predict kidney failure at incremental time horizons using time-updated data. But still, measures of change in these laboratory measurements were not accounted for. Similar features to the ones that will be described shortly have appeared in tangential research [73, 74]. However, these works did not explore whether these features benefited model performance. This section of the thesis takes a clear and simplified look at the measures of change (or *trend features*) that were incorporated into the predictive model and their impact on performance.

3.5.1 Measures of Change

Several measures of change were developed for the final model that was derived and will be presented in **Chapter 5**. For each laboratory measure included in a model, a suite of features was synthesized, each characterizing some intuitive dynamic property of the series.

If X is the set of original feature matrices for the N patients in the dataset,

$$X = \left\{ \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(N)} \right\},$$

where

$$\boldsymbol{X}^{(i)} = \begin{bmatrix} x_{1,1}^{(i)} & \cdots & x_{1,J}^{(i)} \\ \vdots & \ddots & \vdots \\ x_{k_i,1}^{(i)} & \cdots & x_{k_i,J}^{(i)} \end{bmatrix}.$$

Then the matrix for l trend features of the i-th individual is

$$\boldsymbol{E}^{(i)} = \begin{bmatrix} e_{1,1}^{(i)} & \cdots & e_{1,(J\times l)}^{(i)} \\ \vdots & \ddots & \vdots \\ e_{k_i,1}^{(i)} & \cdots & e_{k_i,(J\times l)}^{(i)} \end{bmatrix},$$

where $e_{k_i,(J\times l)}^{(i)}$ is the value of the l-th trend feature at the k-th timepoint in the patient's time series, computed from the j-th feature from the original feature matrix.

If l represents some specific dynamic property of the j-th series of laboratory measurements, for example, a moving maximum of a feature j, then

$$\boldsymbol{e}_{\max(j)}^{(i)}$$

is the $k_i \times 1$ feature vector containing the moving maximum of feature j of patient i, and

$$e_{k,\max(j)}^{(i)}$$

is its value at the k-th timepoint.

Table 3-7 contains the complete set of features that were incorporated into the models. An important property may be observed from the tabulated equations: at each timepoint in a series, only information from previous timepoints is incorporated into the calculation for the k-th value. To incorporate future feature values into the calculation for a past value would constitute a data leak and would render the feature unusable in a clinical setting, as feature values need to be computed using only the presently available patient data. The current formulation for the equations mimics the procedure that would be used to compute these features in a real-world deployment setting.

Table 3-9 tabulates example trend features for a patient's serum creatinine over time. The patient is the same example patient used in the illustrations up to this point. The patient had a complete set of creatinine measurements, requiring no imputation. In the event of missing data, imputation is performed according to the methods in **Section 3.4.1,** prior to constructing the suite of trend features. Then, at each timepoint t_k, $x_{k,j}^{(i)}$ is computed and recorded. Note that time is counting down towards the time of event, as has been the convention in the figures and annotations used up to this point; the equations can be reformulated to accommodate a different time representation.

3.5.2 Analysis

Analysis of the computed measures of change involved adding them in increments to a predictive model and comparing performance at each step. These results are covered in **Chapter 5**.

The cross-correlation between trend features was analyzed. The results for the two laboratory measurements that were eventually included in the final model (**Chapter**

5) are shown in **Figure 3-8 A-B**. This analysis was performed on all of the follow-up visits for all of the patients in the TOH dataset.

Table 3-7: Description of the trend features measuring change in laboratory measurements.

Description	Function
Maximum and Minimum	
Series maximum.	$e^{(i)}_{k,\max(j)} = \max(\boldsymbol{x}^{(i)}_{:k,j})$
Series minimum.	$e^{(i)}_{k,\min(j)} = \min(\boldsymbol{x}^{(i)}_{:k,j})$
Change from series maximum.	$e^{(i)}_{k,\Delta\max(j)} = e^{(i)}_{k,\max(j)} - x^{(i)}_{k,j}$
Change from series minimum.	$e^{(i)}_{k,\Delta\min(j)} = e^{(i)}_{k,\min(j)} - x^{(i)}_{k,j}$
Change from series maximum per unit of elapsed time.	$e^{(i)}_{k,\Delta t\max(j)} = \begin{cases} e^{(i)}_{k,\Delta\max(j)} & k=1 \\ \dfrac{e^{(i)}_{k,\Delta\max(j)}}{t_{\max} - t_k} & k>1 \end{cases}$ $t_{\max}(i,j,k) = \arg\max_{1\le t\le k} x^{(i)}_{t,j}$
Change from series minimum per unit of elapsed time.	$e^{(i)}_{k,\Delta t\min(j)} = \begin{cases} e^{(i)}_{k,\Delta\min(j)} & k=1 \\ \dfrac{e^{(i)}_{k,\Delta\min(j)}}{t_{\min} - t_k} & k>1 \end{cases}$ $t_{\min}(i,j,k) = \arg\min_{1<t<k} x^{(i)}_{t,j}$
Change from Baseline	
Change from series baseline.	$e^{(i)}_{k,\Delta b(j)} = x^{(i)}_{k,j} - x^{(i)}_{1,j}$
Change from series baseline per unit of elapsed time.	$e^{(i)}_{k,\Delta tb(j)} = \begin{cases} 0 & k=1 \\ \dfrac{x^{(i)}_{k,j} - x^{(i)}_{1,j}}{t_1 - t_k} & k>1 \end{cases}$
Tendency	
Series mean.	$e^{(i)}_{k,\mathrm{mean}(j)} = \dfrac{1}{k}\sum_{t=1}^{k} x^{(i)}_{t,j}$

Table 3-7: continued.

Last-3 visit mean.	$e^{(i)}_{k,\text{mean3}(j)} = \frac{1}{\min(3,k)} \sum_{t=\max(1,k-2)}^{k} x^{(i)}_{t,j}$
Recent Rates of Change	
First difference; pairwise change; velocity between two consecutive visits.	$e^{(i)}_{k,\Delta\text{tp}(j)} = \begin{cases} 0 & k = 1 \\ \frac{x_{k,j} - x_{k-1,j}}{t_{k-1} - t_k} & k > 1 \end{cases}$
Mean of the first differences; average velocity.	$e^{(i)}_{k,\text{mean}(\Delta\text{tp}(j))} = \begin{cases} 0 & k = 1 \\ \frac{1}{k-1} \sum_{t=2}^{k} e^{(i)}_{t,\Delta\text{tp}(j)} & k > 1 \end{cases}$
Standard deviation of the first differences; velocity dispersion. [a]	$e^{(i)}_{j,\text{std}(\Delta\text{tp}(j))} = \begin{cases} 0 & k < 3 \\ \sqrt{\frac{1}{k-2} \sum_{t=2}^{k} (e^{(i)}_{t,\Delta\text{tp}(j)} - e^{(i)}_{t,\text{mean}(\Delta\text{tp}(j))})^2} & k \geq 3 \end{cases}$
First difference between two consecutive velocities; acceleration.	$e^{(i)}_{k,\text{a}(j)} = \begin{cases} 0 & k < 3 \\ \frac{e^{(i)}_{k,\Delta\text{tp}(j)} - e^{(i)}_{k-1,\Delta\text{tp}(j)}}{t_{k-1} - t_k} & k \geq 3 \end{cases}$
Mean of the first differences between consecutive velocities; average acceleration.	$e^{(i)}_{k,\text{mean}(\text{a}(j))} = \begin{cases} 0 & k < 3 \\ \frac{1}{k-2} \sum_{t=3}^{k} e^{(i)}_{t,\text{a}(j)} & k \geq 3 \end{cases}$

a: 1) Standard deviation is computed using 1 degree of freedom (ddof). This is an arbitrary and inconsequential decision in the context of these trend features. 2) As with previous trend features that are forced to be imputed with zero at baseline, the standard deviation feature starts to be built at $k = 2$. In other words, the computation is always performed on $k - 1$ time points. The divisor becomes $k - 1 - ddof = k - 2$.

Table 3-8: Example of a set of computed features describing change in patient serum creatinine ($x_{k,j}^{(i)}$) at each time point in the patient's series.

k	t_k	$x_{k,j}^{(i)}$	max	min	Δmax	Δmin	Δtmax	Δtmin	Δb	Δtb	mean	mean3	Δtp	mean (Δtp)	std (Δtp)	a	mean (a)
0	42.8	220	220	220	0	0	0.0	0.0	0	0.0	220.0	220.0	0.0	0.0	0.0	0.0	0.0
1	38.3	280	280	220	0	60	0.0	13.5	60	13.5	250.0	250.0	13.5	13.5	0.0	0.0	0.0
2	34.6	293	293	220	0	73	0.0	8.9	73	8.9	264.3	264.3	3.5	8.5	7.1	-2.7	-2.7
3	30.9	327	327	220	0	107	0.0	9.0	107	9.0	280.0	300.0	9.1	8.7	5.0	1.51	-0.6
4	28.3	252	327	220	-75	32	-29.2	2.2	32	2.2	274.4	290.7	-29.2	-0.8	19.4	-14.9	-5.4
5	23.0	306	327	220	-21	86	-2.7	4.3	86	4.3	279.7	295.0	10.1	1.4	17.5	7.3	-2.2
6	17.6	333	333	220	0	113	0.0	4.5	113	4.5	287.2	297.0	5.0	2.0	15.7	-0.9	-1.9
7	13.2	529	529	220	0	309	0.0	10.4	309	10.4	317.5	389.3	44.2	8.0	21.5	8.8	-0.2
8	11.8	404	529	220	-125	184	-89.3	5.9	184	5.9	327.1	422.0	-89.3	-4.1	39.7	-95.3	-13.8
9	9.2	471	529	220	-58	251	-14.6	7.5	251	7.5	341.5	468.0	26.1	-0.8	38.5	45.0	-6.4
10	5.2	579	579	220	0	359	0.0	9.6	359	9.6	363.1	484.7	27.2	2.0	37.4	0.3	-5.7
11	2.7	838	838	220	0	618	0.0	15.4	618	15.4	402.7	629.3	100.9	11.0	46.3	28.7	-2.2

Abbreviations: k denotes the k-th visit in the patient series occurring at a time t_k before an event; $x_{k,j}^{(i)}$ is the vector of original feature values (creatinine); each of the monikers in the columns that follow represent one of the distinct measures of change defined in **Table 3-7**.

This analysis sought to confirm that each trend feature had the potential to contribute additional (independent) information to a model; regardless of whether these trend features would actually improve model performance.

To demonstrate that the trend features were in fact informative of the outcome (kidney failure), the sample of patients that experienced kidney failure was analyzed with respect to these features. The sample was constructed by taking all the patients who had a kidney failure event and at least 6 months of follow-up time. Patient feature values were selected from the first visit as of this 6-month baseline (if it existed). The patients were then stratified across the mean value for that feature and pooled into a low (L) and a high (H) group. Median survival times in these two groups were then compared, as tabulated in **Table 3-9** and **Table 3-10** for each of the features that were ultimately included in the final model.

3.5.3 Discussion

The correlation analysis revealed that the trend features correlated at varying degrees amongst each other, and to the original feature from which the measures were computed. However, most features exhibited a Spearman correlation $R<0.50$. What this indicates is that these features, should they be informative of the outcome, may in fact contribute supplementary information to the model [75]. On the other hand, some features correlate strongly $R>0.9$, indicating that they may be redundant and not all of them need be added to the model. These trend features include the mean and the max of the patient series.

Correlation between feature values and the outcome of interest (kidney failure) was assessed in **Table 3-9** and **Table 3-10**. The results vary depending on the laboratory measurement and the measure of change. Creatinine showed the greatest difference in median time to kidney failure between the low and high groups. While not a conclusive statement on the potential utility of these features to a model, these indicators (substantial difference in time to

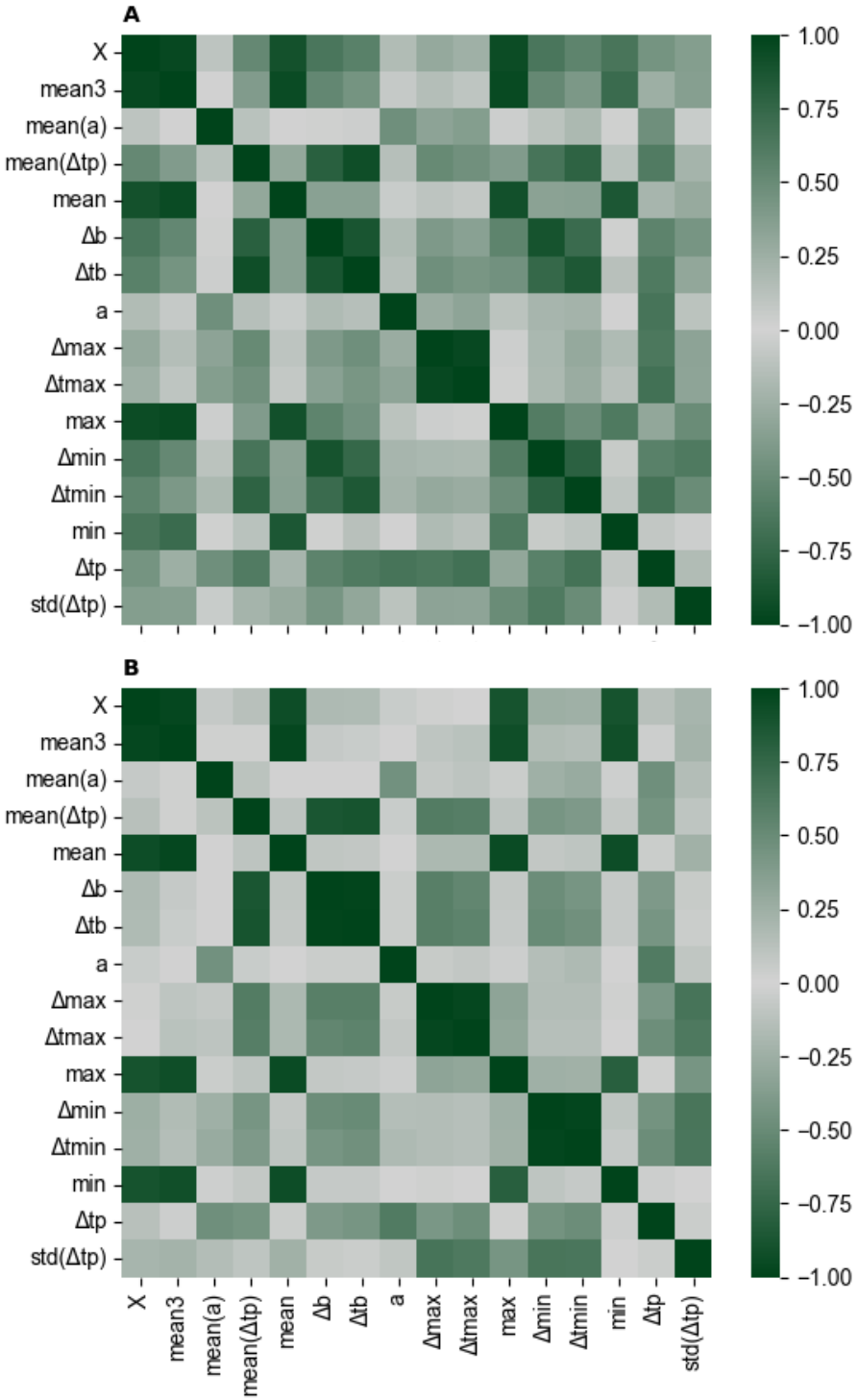


Figure 3-8: Correlation (Spearman's R) between calculated engineered features in (**A**) creatinine, and (**B**) urine albumin-to-creatinine ratio obtained on the complete dataset.

Table 3-9: Median months before kidney failure event for follow-up visits below (L) feature mean and above (H) feature mean (measures of change in serum creatinine).

	Feature Mean	Feature Mean (L)	Feature Mean (H)	Median Months Before Kidney Failure (L)	Months Difference (L - H)
X	365.9	284.9	482.7	19.7	-13.2
mean3	337.9	272.0	425.7	19.7	-12.9
mean(a)	2.0	-3.5	14.2	14.3	-4.2
mean(Δtp)	7.4	0.2	18.9	17.7	-10.5
mean	332.7	270.0	415.0	19.9	-12.9
Δb	59.9	8.6	138.0	18.5	-11.9
Δtb	7.8	1.2	18.5	18.5	-12.0
a	0.9	-5.6	8.9	14.2	-2.8
Δmax	-12.3	-43.8	-0.8	18.1	-7.3
Δtmax	-3.0	-12.1	-0.2	17.8	-6.5
max	378.2	299.4	490.1	19.5	-12.8
Δmin	72.4	26.5	150.3	18.7	-12.7
Δtmin	10.9	3.9	22.1	18.7	-12.5
min	293.5	237.8	365	19.2	-11.2
Δtp	9.4	-2.1	27.5	18.1	-11.0
std(Δtp)	15.4	6.0	33.0	15.6	-6.5

Table 3-10: Median months before kidney failure event for follow-up visits below (L) feature mean and above (H) feature mean (measures of change in urine albumin-to-creatinine ratio).

	Feature Mean	Feature Mean (L)	Feature Mean (H)	Median Months Before Kidney Failure (L)	Months Difference (L - H)
X	233.4	97.7	453.8	15.8	-6.8
mean3	235.9	103.6	444.4	15.8	-6.9
mean(a)	-2.5	-48.2	4.9	9.2	4.1
mean(Δtp)	-3.9	-26.6	4.6	11.8	1.4
mean	238.9	105.2	448.0	15.8	-6.9
Δb	-20.1	-152.8	33.5	12.9	0.2
Δtb	-2.8	-20.3	4.4	12.9	0.2
a	5.0	-2.9	39.8	13.2	-1.7
Δmax	-50.8	-181.8	-4.0	11.5	1.7
Δtmax	-8.1	-30.2	-0.7	11.3	2.0
max	284.1	124.8	534.3	15.6	-6.7
Δmin	32.5	1.7	143.8	13.3	-2.7
Δtmin	6.1	0.4	28.1	13.4	-2.9
min	200.8	82.7	396.0	15.8	-6.9
Δtp	1.7	-7.2	34.0	13.1	-0.2
std(Δtp)	19.7	3.2	63.1	13.9	-3.5

Abbreviations: L, lower group; H, higher group; each of the monikers in table index represent one of the distinct measures of change defined in **Table 3-7**.

kidney failure between the two groups) show that the synthesized features are in fact informative of the outcome.

The remaining results of the feature engineering analysis will be presented in **Chapter 5**, specifically **Appendix B.1, Table B-6**. It will be shown that the explicit incorporation of the described measures of change can in most cases incrementally boost model performance, and in cases where the number of available laboratory measurements are reduced, significantly boost model performance across several performance metrics. In summary, advanced CKD represents a compelling example where the explicit integration of this feature engineering approach into a clinical prediction model could significantly improve prediction accuracy, and thus patient-tailored care. Such analysis is lacking in the literature, and **Chapter 5** provides concrete evidence to support their future use in the clinical use case discussed in this thesis.

3.6 Modeling

The dataset, after the procedures in prior sections have been performed, may be diagrammatically illustrated in the manner of **Figure 3-9**. Survival models, most traditionally the Cox regression model, use only baseline covariate data on each patient, $\boldsymbol{X}_{1,:}^{(i)}$, when estimating parameters using the methods described in **Section 2.2.5**. The traditional Cox regression model, time-varying Cox regression model, and a random survival forest are all survival models which are compared amongst each other and to a random classification forest in **Chapter 4**. Both the traditional (baseline) Cox regression model and the random survival forest utilize only baseline covariate data upon fitting, while the time-varying Cox regression model utilizes the complete set of original feature matrices – $\boldsymbol{X}^{(i)}$ for every i. When predicting survival probabilities, each of these survival models can be applied to each clinic visit, k, using the most-recently available feature data along with the methods described in **Section 2.2**.

The random forest classification model in **Chapter 5** also makes use of the trend features described in **Section 3.5**, denoted $\boldsymbol{E}$ in **Figure 3-9**. At a timepoint, k, this feature vector is computed from the original matrix of features available at timepoint k.

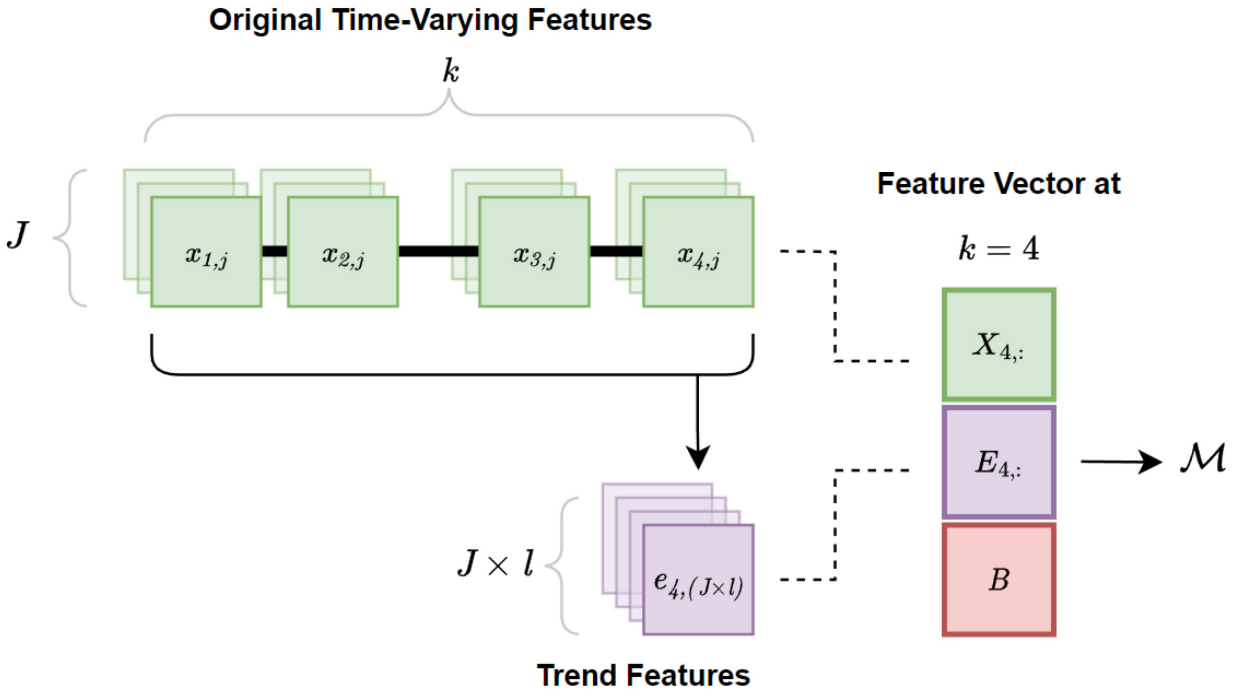


Figure 3-9: Illustration of feature pipeline.

Each of these feature vectors can represent a separate data instance, or example. In simple terms, this example represents an individual patient visit to the clinic nephrologist, with the most recently available clinical data anchored to that timepoint. We evaluate each of the models described based on their ability to accurately classify those visits that are within T months of a kidney failure event, where T is a specified timeframe. Here, and throughout this thesis, kidney failure is defined as the initiation of dialysis or kidney transplantation, where the initiation of dialysis encompasses both those patients that began in a planned manner, and those that began in an unplanned manner. Furthermore, in **Chapter 4** and **Chapter 5** the specific timeframes that are evaluated are 6, 12, and 24 months, with the latter two timeframes being of greatest relevance to this clinical problem.

3.7 Conclusion

The Ottawa Hospital's Multi-Care Kidney Clinic dataset is a unique dataset that offers valuable insights into kidney care outcomes and provides the foundation for this thesis. Nevertheless, careful consideration of limitations is warranted to allow for a balanced interpretation of the results. One important limitation is the absence of datapoints. However, with careful consideration of prior clinical data (where it exists) the impact of missing data can be mitigated to a large extent through imputations. While, nearly all of the current kidney failure risk prediction models are developed and validated to predict longitudinally from the patient's initial (baseline) visit [9, 40, 67], we propose to leverage changes in patient measures over time of follow-up to fine tune the model predictions as more data accumulates for the patient. The nature of patient follow-up in this clinic provides an opportunity to develop and evaluate a model that can operate dynamically at each patient follow-up visit. With methods towards maximizing the quality of existing data, the next chapter evaluates several prediction models that can analyze these data.

Chapter 4: Comparison of Cox Regression and Machine Learning for Short Timeframe Prediction of Kidney Failure among Advanced CKD Patients

4.1 Preamble

This fourth chapter is a modified copy of the manuscript entitled *Short Timeframe Prediction of Kidney Failure among Patients with Advanced Chronic Kidney Disease*, published in *Clinical Chemistry*, 2023 [In Press]. Modifications were made to improve the flow within the context of this thesis.

This manuscript compares Cox regression models to machine learning models in the prediction of kidney failure at timeframes of 6, 12, and 24 months. This analysis sought to answer which out of a reasonable selection of traditional or machine learning models was the best-suited and most effective solution to the clinical problem of short timeframe kidney failure prediction.

Authors: Martin M. Klamrowski, Ran Klein, Christopher McCudden, James R. Green, Tim Ramsay, Babak Rashidi, Christine A. White, Matthew J. Oliver, Ayub Akbari, and Gregory L. Hundemer

Contributions: This article was a collaborative effort between all of the authors listed above. MMK set up the experiment, performed all analyses, prepared the first draft, and contributed major revisions to the final article. GLH, RK, had a substantial influence on the writing of the introduction and discussion sections as well as the overall preparation, presentation, and review of the manuscript. CM had a substantial influence on the compilation and writing of supplementary

materials. MMK, RK, CM, JRG, TR, BR, CAW, MJO, AA, GLH all contributed substantial machine learning, statistical, and/or clinical expertise.

4.2 Methods

4.2.1 Study Design

We performed a retrospective cohort study of adults (≥18 years) with advanced CKD referred to the Ottawa Hospital Multi-Care Kidney Clinic. We compared the performance of internally-derived kidney failure risk prediction models over 6-, 12-, and 24-month timeframes using traditional Cox regression models, and machine learning algorithms that incorporated baseline data alone as well as time-updated data. Cox regression models included: a) Cox regression using baseline variables and b) Cox regression with time-varying variables. Machine learning algorithms included: c) random survival forest and d) random forest classifier.

4.2.2 Study Population

The study population was derived from a database of all patients referred to the clinic from January 1, 2010 with follow-up data available through May 31, 2021 as described in **Chapter 3**. Patients were excluded from the study if they were <18 years of age, selected conservative management, were lost to follow-up, had missing predictor values, or did not develop kidney failure or death and had <24 months of follow-up. The final exclusion criterion was included due to inability to label the outcomes for these patients in all studied models.

The clinical and research activities being reported are consistent with the Principles of the Declaration of Helsinki. All protocols were approved by the Ottawa Health Science Network Research Ethics Board (Protocol ID #20150457-01H). Informed consent requirements were waived due to the retrospective nature of the data. The reporting of this study follows guidelines

for transparent reporting of artificial intelligence algorithms in medicine (MI-CLAIM Checklist, **Supplemental Table A-1**) [76].

4.2.3 Outcomes

The outcome of interest was kidney failure, defined as dialysis or kidney transplantation. Unplanned dialysis was defined as initiating dialysis in the inpatient setting. Planned dialysis was defined as initiating dialysis in the outpatient setting or undergoing pre-emptive kidney transplantation.

4.2.4 Model Development

As previously mentioned, we derived and compared the performance of two Cox regression models using [a) baseline variables, b) time-varying variables] and two machine learning models [c) random survival forest, d) random forest classifier] in the prediction of kidney failure over 6, 12, and 24 months. We assessed the performance of each model using increasingly larger variable sets (**Figure 4-1**). Algorithm and model details are available in the **Supplemental Methods** (**Appendix A.1**).

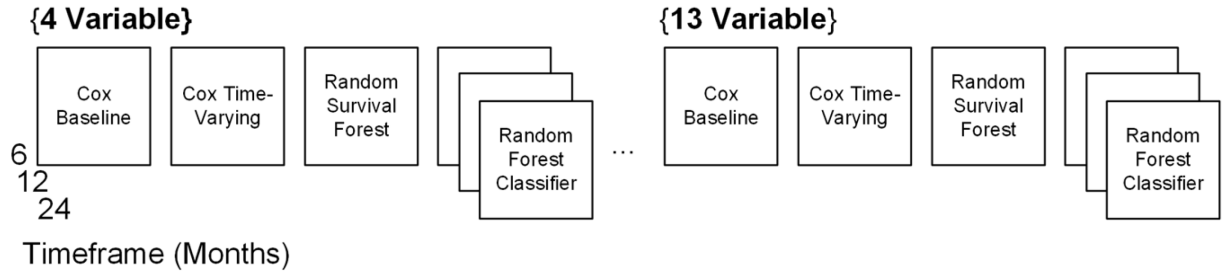


Figure 4-1: Illustration of experiment design.

4.2.5 Variables

Variables included in the study models were selected *a priori* from variables incorporated in pre-existing kidney failure risk prediction models including the KFRE, Grams, and VA models [10, 11, 49, 77]. We used 4-variable, 8-variable, 10-variable, and 13-variable sets within each of our study models. The 4-variable set included age, sex, eGFR, and urine albumin-to-creatinine ratio (ACR) so as to align with the 4-variable KFRE [10, 49]. The larger variable sets iteratively incorporated variables from the aforementioned models (**Table 4-1**). To characterize each patient visit, variable values were assigned based on the closest available measurements, where possible.

Fitting of baseline Cox and random survival forest models used only the baseline visit data, whereas fitting of the time-varying Cox and random forest classifier models used data from all visits. All models were validated using the same data – the most recent clinical data at each patient visit (as per the intended clinical use of the models). Furthermore, while the time-varying Cox and random forest classifier models incorporated updated clinical data into the fitting process, none of the models in this study incorporated changes in variable values. That is to say, when predicting the probability of kidney failure at each visit, each of the models used only the variable values tied to that visit. No historical trends in that patient's variables were incorporated.

4.2.6 Statistical Analysis to Compare Model Performance

Each model produced a predicted probability of kidney failure, which was then compared to a probability cutoff to classify whether kidney failure is expected or not. Each model was evaluated based on its ability to classify whether kidney failure occurred within 6, 12, or 24 months of a clinic visit.

Five-fold cross-validation with stratified cold-patient splits was performed, meaning every fold contained a class-balance representative of the overall data, and each fold did not contain visits from patients in other folds. Models were compared using metrics of area under the receiver

operating characteristic curve (AUC-ROC) and area under the precision-recall curve (AUC-PR) to summarize discriminative ability at each probability cutoff. Precision-recall curves plot precision versus sensitivity, which provides a better measure of predictive performance in the case of imbalanced datasets, such as is the case here (i.e., more visits without outcome event than with outcome event within the timeframe in question) [75]. Brier scores were computed to summarize model calibration and discrimination. We used the maximum precision obtainable at 70% recall (PrRe70) as our main performance metric. PrRe70 can be considered a closer measure of the potential clinical impact of the models [78], representing concrete stepwise error rates (false positives vs. false negatives) at a meaningful probability cutoff. Performance metrics were calculated as the mean value over cross-validation folds, and 95% confidence intervals (CI) were obtained from 10,000 bootstrap samples of fold performances. Finally, as an external benchmark, the 4-variable KFRE model [10, 49] was evaluated at its intended timeframe of 24 months.

4.2.7 Model Examinations

We performed two tiers of analyses to examine the machine learning models [76]. Variable permutation importance on Brier score was done to gauge model sensitivity when specific variables were randomly de-correlated with the outcome variable. Shapley Values [79, 80] were computed to summarize the direction and degree of specific variable contributions to model output.

To further examine the behavior of each model type, we plotted a curve representing the average predicted probability of kidney failure for all patients who had a kidney failure event at each follow-up prior to their dialysis start. We stratified these patients into unplanned and planned dialysis.

Additional examination insights are implicit in the experiment design (cross-validation, increasing number of variables). All analyses were performed in Python (v3.10) using open-source libraries (**Supplemental Table A-2**).

4.2.8 External Testing

The top performing models were tested in two external settings to assess generalizability: Kingston General Hospital (Kingston, Ontario, Canada) and University Health Network (Toronto, Ontario, Canada). Data from these two sites were combined to create a single external validation set. For each of the studied algorithms, we selected the variable set that yielded the highest 12-month PrRe70 in validation, for succinctness. Final models were trained on the entire Ottawa derivation cohort and then tested on the combined external datasets using the performance metrics previously described. Data processing and labeling procedures were repeated for these external cohorts. Means, 2.5 and 97.5 percentiles were extracted from 1,000 bootstrap resamples of the external patient sets.

4.3 Results

4.3.1 Cohort Description

From a total of 2,432 consecutive patients with advanced CKD referred to the Ottawa Hospital Multi-Care Kidney Clinic, 1,757 met our inclusion criteria. Reasons for exclusion were as follows: age <18 years (n=2), conservative management selected by patient (n=117), loss to follow-up (n=193), missing predictor values (n=276), and patient did not develop kidney failure or death and had <24 months of follow-up (n=87). The baseline characteristics of the study cohort are summarized in **Table 4-1**. Within the study window, 1,204 (69%) of patients developed kidney failure, 347 (20%) died prior to kidney failure, and 206 (12%) were still being followed in the clinic.

4.3.2 Comparison of Model Performance

Performance results are tabulated in **Table 4-2** for all model types, variable sets, and timeframes. In general, performance metrics changed very little by increasing the number of variables included in each model beyond the initial 4-variable model. To compare the performance of each model, **Figure 4-2** displays the performance metrics of the best model (i.e., number of variables) for each model type over each timeframe. Compared to the baseline Cox model, the machine learning models and time-varying Cox model had higher performance metrics in the 6-month timeframe by a statistically significant margin. These trends persisted but were less pronounced in the 12-month prediction models, not reaching statistical significance (overlapping 95% CI). With the 24-month prediction models, the performance was similar across all models.

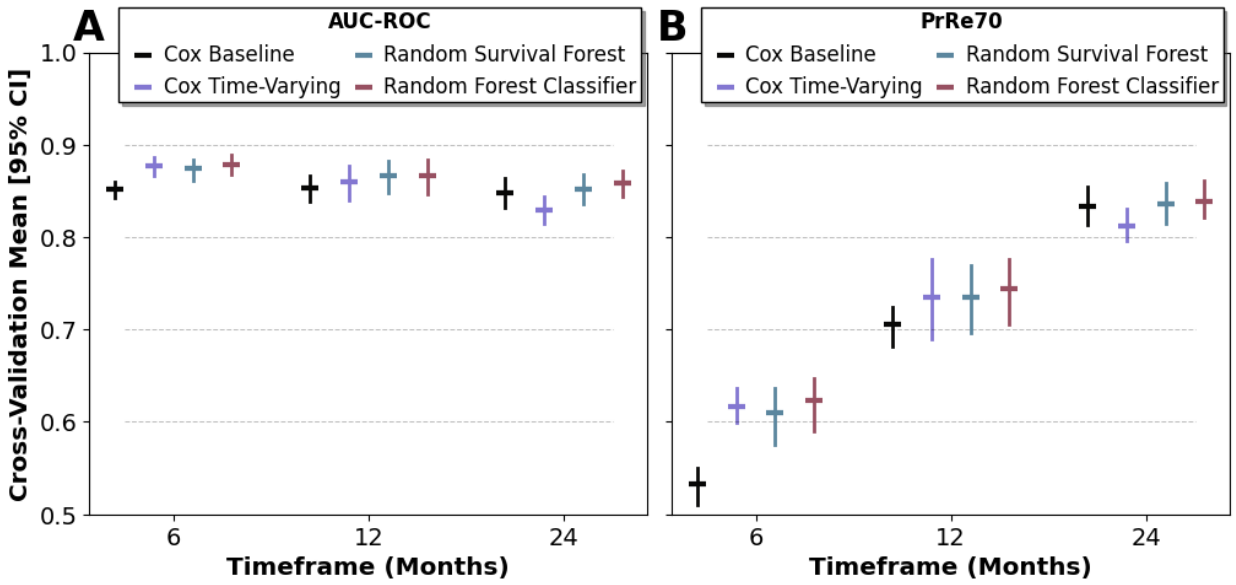


Figure 4-2: Comparison of model performance. (**A**) Area under receiver-operating characteristic curve (AUC-ROC), (**B**) maximum precision at a recall of 70% (PrRe70), plotted for each model type and timeframe. Performances are taken from bolded elements in **Table 4-2** representing each model type with the variable set that yielded its highest performance based upon PrRe70.

Notably, the random forest classifier showed the highest performance in each timeframe. However, as stated above, this difference was greatest at the shortest timeframe (6 months) and marginal at the longest timeframe (24 months). AUC-PR scores yielded similar patterns

(**Supplemental Table A-3**). While Brier scores were similar across all models and variable sets (in the 0.11-0.21 range), they tended to increase over longer timeframes as class imbalance lessened (**Supplemental Table A-4**).

Table 4-1: Baseline characteristics of derivation cohort (N = 1757).

Variable	Summary Statistics	Inclusion in Variable Sets			
		4 Variable	8 Variable	10 Variable	13 Variable
Demographics					
Age, Years, Mean (SD)	66 (15)	X	X	X	X
Male Sex, N (%)	1104 (63)	X	X	X	X
Laboratory Data[a]					
Creatinine, mg/dL, Mean (SD)	3.49 (1.10)				
eGFR, mL/min/1.73m^2, Mean (SD)	18 (6)	X	X	X	X
Urine Albumin-to-Creatinine Ratio, mg/g, Median (IQR)	1450 (424, 2953)	X	X	X	X
Calcium, mg/dL, Mean (SD)	8.94 (0.60)		X	X	X
Phosphate, mg/dL, Mean (SD)	4.24 (0.96)		X	X	X
Bicarbonate, mEq/L, Mean (SD)	24 (3)		X	X	X
Albumin, g/dL, Mean (SD)	3.5 (0.5)		X	X	X
Potassium, mEq/L, Mean (SD)	4.5 (0.6)				X
Comorbidities, N (%)					
Diabetes Mellitus	1054 (60)			X	X
Hypertension	1598 (91)			X	X
Congestive Heart Failure	386 (22)				X
Medication Use, N (%)					
ACE Inhibitor/ARB	771 (44)				
Diuretic	1005 (57)				
Vital Signs/Anthropometrics					
Systolic Blood Pressure, mmHg, Mean (SD)	137 (21)				X
Diastolic Blood Pressure, mmHg, Mean (SD)	72 (13)				
Body Mass Index, kg/m^2, Mean (SD)	30.0 (7.1)				

Abbreviations: ACE, angiotensin-converting-enzyme; ARB, angiotensin II receptor blocker; eGFR, estimated glomerular filtration rate; IQR, interquartile range; N, number; SD, standard deviation.

[a] Laboratory data presented in traditional units. Conversion to International System of Units (SI) is as follows: creatinine conversion factor (CF) 88.42 to convert to µmol/L; urine albumin-to-creatinine ratio CF 0.113 to convert to mg/mmol; calcium CF 0.2495 to convert to mmol/L; phosphate CF 0.3229 to convert to mmol/L; bicarbonate CF 1.0 to convert to mmol/L; albumin CF 10.0 to convert to g/L; potassium CF 1.0 to convert to mmol/L.

4.3.3 Comparison of 24 Month Model Performance to the Kidney Failure Risk Equation

Our internally-derived 4-variable baseline Cox model had similar performance to that of the 4-variable KFRE (AUC-ROC 0.85 [95%CI 0.83-0.87] vs. 0.84 [95%CI 0.82-0.86], PrRe70 0.83 [95%CI 0.80-0.85] vs. 0.82 [95%CI 0.79-0.84]) at the 24-month timeframe within our study cohort.

Table 4-2: Cross-validation performance results of 6-, 12-, and 24-month models across variable sets in derivation cohort.

		6-Month		12-Month		24-Month	
		AUC-ROC (95% CI)	**PrRe70 (95% CI)**	**AUC-ROC (95% CI)**	**PrRe70 (95% CI)**	**AUC-ROC (95% CI)**	**PrRe70 (95% CI)**
Cox Baseline	*4 Variable*	0.85 (0.84, 0.86)	**0.53 (0.51, 0.55)**	0.85 (0.84, 0.87)	**0.71 (0.68, 0.72)**	0.85 (0.83, 0.87)	0.83 (0.80, 0.85)
	8 Variable	0.85 (0.84, 0.86)	0.53 (0.50, 0.55)	0.85 (0.84, 0.86)	0.70 (0.67, 0.72)	0.85 (0.83, 0.86)	0.83 (0.81, 0.85)
	10 Variable	0.85 (0.84, 0.86)	0.52 (0.50, 0.55)	0.85 (0.83, 0.86)	0.70 (0.68, 0.72)	0.85 (0.83, 0.86)	**0.83 (0.81, 0.85)**
	13 Variable	0.85 (0.84, 0.86)	0.52 (0.50, 0.54)	0.85 (0.84, 0.86)	0.69 (0.67, 0.71)	0.85 (0.83, 0.86)	0.83 (0.80, 0.85)
Cox Time-Varying	*4 Variable*	0.88 (0.87, 0.89)	**0.62 (0.60, 0.64)**	0.86 (0.84, 0.87)	0.73 (0.68, 0.76)	0.83 (0.81, 0.84)	0.81 (0.79, 0.83)
	8 Variable	0.88 (0.87, 0.89)	0.62 (0.59, 0.65)	0.86 (0.84, 0.88)	**0.73 (0.69, 0.77)**	0.83 (0.82, 0.84)	0.81 (0.79, 0.83)
	10 Variable	0.88 (0.87, 0.89)	0.62 (0.59, 0.64)	0.86 (0.84, 0.88)	0.73 (0.69, 0.78)	0.83 (0.81, 0.84)	**0.81 (0.79, 0.83)**
	13 Variable	0.88 (0.87, 0.89)	0.61 (0.58, 0.65)	0.86 (0.84, 0.88)	0.73 (0.69, 0.77)	0.83 (0.81, 0.84)	0.81 (0.79, 0.83)
Random Survival Forest	*4 Variable*	0.87 (0.86, 0.88)	**0.61 (0.57, 0.64)**	0.86 (0.84, 0.88)	0.73 (0.69, 0.76)	0.85 (0.83, 0.86)	0.82 (0.80, 0.85)
	8 Variable	0.88 (0.87, 0.89)	0.61 (0.58, 0.63)	0.87 (0.85, 0.88)	**0.74 (0.70, 0.77)**	0.85 (0.83, 0.87)	0.83 (0.81, 0.85)
	10 Variable	0.88 (0.87, 0.89)	0.60 (0.57, 0.63)	0.87 (0.85, 0.88)	0.73 (0.69, 0.76)	0.85 (0.83, 0.87)	**0.84 (0.82, 0.86)**
	13 Variable	0.88 (0.87, 0.89)	0.60 (0.56, 0.63)	0.87 (0.85, 0.88)	0.73 (0.69, 0.76)	0.85 (0.83, 0.87)	0.83 (0.81, 0.85)
Random Forest Classifier	*4 Variable*	0.87 (0.86, 0.88)	0.61 (0.58, 0.64)	0.86 (0.84, 0.87)	0.73 (0.70, 0.76)	0.85 (0.83, 0.86)	0.83 (0.81, 0.84)
	8 Variable	0.88 (0.87, 0.89)	0.62 (0.59, 0.65)	0.87 (0.85, 0.88)	**0.74 (0.71, 0.78)**	0.86 (0.84, 0.87)	0.84 (0.82, 0.86)
	10 Variable	0.88 (0.87, 0.89)	**0.62 (0.59, 0.65)**	0.87 (0.85, 0.88)	0.74 (0.70, 0.78)	0.86 (0.84, 0.87)	0.84 (0.82, 0.86)
	13 Variable	0.88 (0.87, 0.89)	0.62 (0.59, 0.64)	0.87 (0.85, 0.88)	0.74 (0.70, 0.78)	0.86 (0.84, 0.87)	**0.84 (0.82, 0.86)**

Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; CI, confidence interval; PrRe70, maximum precision at 70% recall.
For each model / timeframe pair, the single highest performing model (amongst number of variables) is bolded based upon PrRe70. For example, the time-varying Cox model obtained its best 24-month-PrRe70 with 10 variables. The

baseline Cox model also obtained its maximal PrRe70 at 24 months with 10 variables. On the other hand, the random forest classifier achieved its maximal PrRe70 at 24 months with 13 variables.

Further, the variables most strongly associated with the outcome of kidney failure for all study models matched with those incorporated into the well-validated KFRE (**Supplemental Tables A5-8, Supplemental Figures A1-2**) [10, 49].

4.3.4 Model Examinations

Shapley analysis revealed machine learning model variable importance was correlated with Cox regression hazard-ratios (**Supplemental Tables A5-6, Supplemental Figures A1-2**). Brier permutation scores show the models were heavily reliant on eGFR and urine ACR and were not significantly affected by the other included variables (**Supplemental Tables A7-8**). For the unplanned dialysis subgroup, machine learning models were superior on average to Cox models throughout the full 6–12-month period prior to kidney failure (**Figure 4-3**).

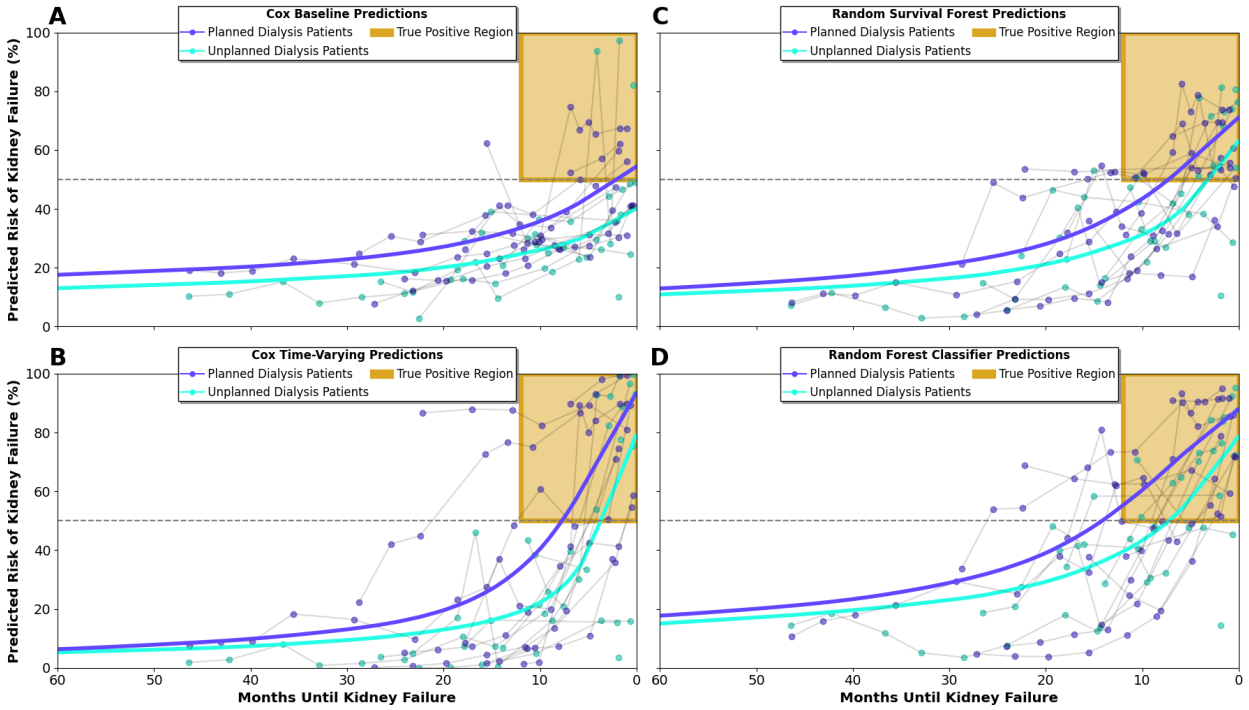


Figure 4-3: Predicted kidney failure risk across longitudinal follow-up among randomly sampled patients by study model. For each model, the 12-month predictions are plotted for the same 10 randomly sampled planned dialysis

patients (purple) and the same 10 randomly sampled unplanned dialysis patients (cyan), obtained over the five-fold cross-validation procedure. Each line represents the course of one patient over the study, with individual visits denoted by marks. The true positive region is highlighted in gold, representing the approximate interval in which a model should "fire" in order to catch patients in need of dialysis in a timely and precise manner. The horizontal dashed line represents a 50% probability cutoff threshold. Any marks (visits) above the dashed line and outside the gold region represent false positives. Visits under the gold region represent false negatives. A local polynomial regression (LOESS) line of best-fit shows the average probability output by the model at each time prior to kidney failure. All planned and unplanned patients were used to fit each respective curve for the subgroups. Separate panels are plotted for (**A**) the Cox baseline model, (**B**) time-varying Cox, (**C**) the random survival forest, (**D**) and the random forest classifier.

Table 4-3: External testing performance results of selected models across 6, 12, and 24 month timeframes in external validation cohort.

		6-Month		12-Month		24-Month	
		AUC-ROC (95% CI)	**PrRe70 (95% CI)**	**AUC-ROC (95% CI)**	**PrRe70 (95% CI)**	**AUC-ROC (95% CI)**	**PrRe70 (95% CI)**
Cox Baseline	*4 Variable*	0.85 (0.84, 0.87)	0.60 (0.56, 0.65)	0.86 (0.84, 0.88)	0.78 (0.75, 0.82)	0.86 (0.84, 0.88)	**0.88 (0.85, 0.91)**
Cox Time-Varying	*8 Variable*	0.86 (0.84, 0.87)	0.63 (0.59, 0.67)	0.84 (0.82, 0.86)	0.76 (0.71, 0.81)	0.82 (0.79, 0.84)	0.82 (0.78, 0.86)
Random Survival Forest	*8 Variable*	0.85 (0.83, 0.87)	0.61 (0.57, 0.65)	0.84 (0.82, 0.87)	0.77 (0.73, 0.81)	0.84 (0.81, 0.86)	0.85 (0.81, 0.88)
Random Forest Classifier	*8 Variable*	0.85 (0.84, 0.87)	**0.64 (0.60, 0.68)**	0.85 (0.83, 0.87)	**0.79 (0.74, 0.83)**	0.85 (0.82, 0.87)	0.86 (0.82, 0.90)

Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; CI, confidence interval; AUC-PR, area under the precision-recall curve; PrRe70, maximum precision at 70% recall.
The highest performing model at each timeframe (based upon PrRe70) is bolded.

4.3.5 External Testing

Data from two independent advanced CKD cohorts from Kingston General Hospital (n=493) and University Health Network (n=209) were combined to create a single cohort for external validation testing (**Supplemental Table A-9**). External testing performance metrics are tabulated in **Table 4-3** and **Supplemental Tables A10-11**. No significant decay in performance was observed for any of the models. The random forest classifier remained the best performing model at 6 and 12

months based upon PrRe70 while the baseline Cox model was the highest performing model at 24 months (**Table 4-3**). The random forest classifier was the most probabilistically accurate model (Brier score) at the 6- and 12-month timeframes (**Supplemental Table A-11**). Overall, the trends in all metrics did not fluctuate significantly from internal validation results. Increases between internal results and external results can be in part attributed to increased positive class prevalence in the external test set.

4.4 Discussion

To the best of our knowledge, this is the first study to present a model explicitly tied to the clinical purpose of preventing unplanned dialysis by predicting short timeframe (6 and 12 months) kidney failure risk among patients with advanced CKD using commonly available laboratory and demographic data. In this retrospective cohort study of 1,757 consecutive patients with advanced CKD, we found that machine learning models outperformed traditional Cox regression models in predicting kidney failure risk over these short timeframes. Overall, the random forest classifier consistently showed the highest performance at 6- and 12-month timeframes. When the timeframe for prediction of kidney failure was extended to 24 months, performance was similar across all models. All selected models generalized to two independent external cohorts, indicating the potential for their widespread application for short timeframe prediction of kidney failure risk.

Our primary interest was in predicting kidney failure among patients with advanced CKD over shorter timeframes (6 and 12 months) than those for which traditional risk prediction models are designed. Shorter-term predictions may be particularly useful in this high-risk population to better express the urgency with which the planning for kidney failure (e.g., modality education, access creation, and transplant evaluation) must take place, which the traditional 2–5-year prediction models fail to fully capture. For example, a retrospective study from the United States Renal Data System showed that the optimal timing of arteriovenous fistula (AVF) creation was 6-

9 months pre-dialysis as this allowed adequate time for access maturation while limiting the additional interventional access procedures required to maintain longer-existing AVFs [81]. A model tailored to predict kidney failure risk over a shorter time horizon would be better suited to match this unmet need. Furthermore, unplanned dialysis remains a major problem with 40-60% of CKD patients who progress to kidney failure initiating dialysis in an unplanned fashion [3-5, 8, 82-88]. Unplanned dialysis is strongly linked to increased patient morbidity and mortality [3], worse patient-reported outcome measures such as mental health and quality of life [89], and a heavy financial burden to healthcare systems [8].

Nephrologists' clinical judgement alone surrounding short-term kidney failure risk has previously been shown to have high sensitivity but poor specificity [90]. Despite the existence of current kidney failure risk prediction models to augment clinical judgement, the high rates of unplanned dialysis suggest an inability of nephrologists to use traditional risk prediction tools to reliably identify and prepare patients at high risk for imminent dialysis, even among advanced CKD patients followed in specialized nephrology clinics over an extended follow-up period [4, 5, 82, 86]. This suggests that these traditional models may not be informative when applied at each follow-up visit, or may not be directly interpretable in relation to the specific clinical decision of whether to prepare a patient for dialysis now or not. Thus, there may be only marginal value added with use of these traditional longer timeframe models in communicating the urgency with which preparation for dialysis needs to occur as compared to having a model that directly addresses the risk for imminent kidney failure. We now demonstrate the superior utility of the machine learning models in this regard (**Figure 4-3**). The curve for the baseline Cox model is flatter than the curves for the other models, making the interpretation of risk difficult. Furthermore, the random forest classifier's mean predicted probability for unplanned dialysis patients is higher throughout the illustrated 12-month window prior to kidney failure while maintaining similar or better overall PrRe70. This means, at a probability cutoff of 50%, out of the unplanned dialysis patients that

were followed for at least 6 months, 44% (95%CI 34-53%) were detected in a timely manner (predicted risk ≥50% 6-9 months prior their kidney failure) by the random forest classifier model as compared to 13% (95%CI 7-20%), 14% (95%CI 7-20), and 27% (95%CI 19-36%) by the baseline Cox, time-varying Cox, and random survival forest models, respectively. Associated false-positive rates were 0.69, 0.42, 0.72, and 0.69, false-positives for every unplanned dialysis patient detected by the random forest classifier, baseline Cox, time-varying Cox, and random survival forest models, respectively. This suggests that the random forest classifier may be better suited to identify and prevent unplanned dialysis by concretely expressing the urgency with which preparation must occur all while maintaining low false-positive rates.

Importantly, all selected models generalized to two independent external advanced CKD cohorts. While differences between selected models generally did not reach statistical significance, the principal result from the derivation cohort was preserved (random forest classifier performs better than baseline Cox at short timeframes of 6 and 12 months). We emphasize that all three cohorts are relatively small by predictive modeling standards, making statistical significance more difficult to obtain. Were any model to be released for more widespread use, it would undoubtedly have to be trained and tested over multiple centers with tens of thousands of patients, as established prediction models have been [10, 11, 91]. Nonetheless, if we appreciate the aforementioned considerations around the results in **Figure 4-3** in tandem with the random forest classifier's improved PrRe70 and improved probabilistic accuracy at 6 and 12 months over the other models both internally and externally, the random forest classifier becomes the best model proposition for short timeframe kidney failure risk prediction.

We acknowledge several limitations within our study. First, these models were derived in a single center, which may not account for differences in patient populations and provider practice (e.g., timing of dialysis initiation) in other centers. Even at this single institution, data is not standardized due to multiple labs employing non-standardized testing, changes in practice over

the course of data acquisition, and a large number of clinicians. This emphasizes the ongoing need for refinement and standardization of laboratory techniques, including how often lab samples are collected for each patient [92]. Second, our external testing sets were relatively small in regard to cohort size. While an important evaluative data point, further testing would be required to conclusively evaluate external performance. Third, we did not explore feature engineering, data augmentation, model tuning, or other machine learning engineering techniques which are likely to enhance the prediction ability of machine learning models [75]. However, we want to emphasize our intent herein was to evaluate the suitability of several predictive algorithms for predicting kidney failure risk over short timeframes using routinely collected clinical laboratory data rather than develop a model for more widespread use. Fourth, future studies will be necessary to determine the feasibility and cost-effectiveness of implementing these models into routine advanced CKD practice. For example, what is the cost-tradeoff between catching more unplanned dialysis starts versus preparing some patients for dialysis earlier than may be necessary? Finally, in reference to the goal of using short timeframe prediction of kidney failure among advanced CKD patients to prevent unplanned dialysis, we acknowledge that rates of unplanned dialysis will never be reduced to zero. This is due to the makeup of unplanned dialysis starts including acute kidney injury, lack of nephrology referrals, and unwillingness of some patients to see nephrologists or participate in the kidney failure preparation process. But given the alarmingly high rates of unplanned dialysis (approximately half of all dialysis starts), even modest reductions via improved short timeframe kidney failure risk prediction will likely translate into substantial benefits to both patients and the healthcare system as a whole.

In summary, short timeframe kidney failure prediction models that can be used in a time-updated manner may serve to augment clinical decision-making by accurately identifying patients at high risk for imminent dialysis, thereby allowing for optimal pre-emptive intervention. Of the

models we examined in this study, the random forest classifier may be best suited for this purpose. These findings will inform the development of future kidney failure risk prediction tools.

Chapter 5: Derivation and Validation of a Machine Learning Model for the Prevention of Unplanned Dialysis among Patients with Advanced CKD

5.1 Preamble

This fifth chapter is a copy of a manuscript entitled *Derivation and Validation of a Machine Learning Model for the Prevention of Unplanned Dialysis among Patients with Advanced CKD*, being prepared for submission to an undetermined venue.

This article directly builds upon the experiments performed in **Chapter 4** by improving upon and tuning the best model from those experiments. The incorporation of additional features measuring trend (**Section 3.5**) provided an opportunity to reduce the number of required laboratory variables to only the most commonly available ones while achieving better overall predictive performance than the 8-variable random forest classifier from **Chapter 4**. What follows is an extensive characterization of this final model by focusing on its external validation, potential reduction in unplanned dialysis starts, sensitivity analysis with respect to laboratory measurements, and performance across a number of strata. As such, this **Chapter 5** represents a comprehensive proposal for new short timeframe kidney failure risk prediction models to reduce the burden of unplanned dialysis in advanced CKD care centers.

5.2 Methods

5.2.1 Study Design and Populations

A retrospective cohort study was carried out in three specialized healthcare centers located in Ontario, Canada – cohorts dedicated to the specialized treatment of advanced CKD patients. This model derivation study focused on these specific populations. Resultant models were built using a representative cohort from The Ottawa Hospital Multi-Care Kidney Clinic in Ottawa, Ontario, Canada (collection period from 2010 to 2021). Analyses focused on deriving and evaluating the models internally, with external validation then being obtained from two other centers in Kingston and Toronto (collection periods 2015-2023) (**Table 3-2**). Reasons for exclusion were as follows: age <18 years, conservative management selected by patient, loss to follow-up, and patient did not develop kidney failure or death and had <12 months of follow-up. The nature of patient care is similar across these centers. At each follow-up visit, patients are provided with a comprehensive care package from a team of healthcare professionals. This team typically includes a nurse, a dietician, and a nephrologist, who evaluate and address patient needs. Supplementary support from a pharmacist and social worker is also available. The frequency of patient follow-ups differs among the centers, but, it is typical practice to arrange regular visits every two weeks to six months.

5.2.2 Variables

<u>Independent Variables</u>

Over the course of each patient's follow-up, laboratory measurements are drawn at routine intervals and then tested at times loosely proximate to the visit times with the clinic MD. We anchored observed laboratory values to each clinic MD visit using a forward-fill approach (LOCF), as detailed in **Section 3.4**. For each patient, this yielded a short (in number of timepoints) and

irregularly sampled time-series with the most recently available clinical characteristics and laboratory measurements tied to each timepoint.

The collected clinical characteristics included age and sex, as well as several commonly obtained laboratory measurements (**Table 3-2**). Trends and descriptions of change in these laboratory measurements provided an opportunity to supply additional variable information and predictive power to the models [93-95] (**Table B-1**). The selected variable set included age, sex, serum creatinine, urinary albumin-to-creatinine ratio (uACR), and associated trends and descriptions of change in the patient's repeated measurements, as detailed in **Section 3.5**.

The variable missingness rate was generally below 5% for the above stated model variables (**Table 3-2**). For any variables with greater missingness, simple statistical procedures to test for associations in the missingness allowed us to characterize any potential biases [69] (**Table B-2**). Imputation of missing variables involved a forward pass and a backward pass through each patient series. Imputation on the forward pass involved carrying the last and most-recently available observation forward (LOCF). Missing baseline variables were backfilled, whereby the next observation was carried backwards (NOCB).

Dependent Variables

Models were built to identify the conditions for imminent kidney failure risk. Kidney failure was defined as kidney replacement therapy (KRT), meaning the initiation of dialysis or kidney transplantation. Initiation of dialysis encompassed both unplanned dialysis (initiation in the inpatient setting), and planned dialysis (initiation in the outpatient setting or undergoing pre-emptive kidney transplantation). The particularity of our modeling strategy involved labeling follow-up visits according to whether they fell within 6 or 12 months of a KRT event. The resultant model would thus be conditioning inputs on the identification of kidney failure within 6 or 12 months of visit to the clinical nephrologist. We assigned visits falling within 6 or 12 months of a

KRT event to the positive class. This created a relative class imbalance, which we sought to mitigate using class weighting. Each class was weighted by the inverse of its prevalence.

5.2.3 Statistical Analyses

Modeling

Random decision forest classification algorithms were used [63, 96]. These supervised learners would take as input the aforementioned independent variables and attempt to assign to each follow-up visit a probability of membership to the defined positive class (KRT within 6 or 12 months). We built two models – one with a 6-month prediction timeframe, and a second with a 12-month prediction timeframe. Optimal model hyperparameters were exhaustively evaluated by grid-searching a set of predefined options (**Table B-3**) [61].

Experiment Design

Internal training and calibration data were split 90% and 10%, respectively. Internal performance evaluations were obtained over 5-fold cross-validation procedures [75, 97]. Patient sets across the folds were ensured to be disjoint, and class ratios within each fold were ensured to be representative of the original data. The final models were trained on the 90% partition, calibrated on the 10% partition, and then externally validated on the complete external dataset comprised of the two external cohorts. Both internally (cross-validation) and externally (test), patient sets and associated predictions were bootstrapped with 1,000 repetitions to obtain distributions around summative statistics. Unless otherwise noted, confidence intervals (CI) throughout the reported analyses represent the 2.5 and 97.5 percentiles (95% CI) over the results of these bootstrapped patient sets.

Discrimination

Discrimination metrics included area under the receiver operating characteristic curve (AUC-ROC) and area under the precision-recall curve (AUC-PR) to summarize discriminative ability at each cut-off threshold. Precision-recall (PR) curves plot the precision versus the sensitivity to identifying the positive label, which are informative measures of predictive performance in the case of imbalanced datasets, such as is the case here (i.e., more visits without outcome event than with outcome event within the 6- or 12-month timeframe) [75].

Calibration

Calibration plots were obtained in both the derivation and validation cohorts. Predictions were grouped by quintiles. Brier scores were computed to summarize the accuracy of the model's probabilistic predictions. In classification contexts, the Brier score is an important measure of overall calibration and probabilistic accuracy [98].

Impact Measures

As in [73, 74], histograms illustrating the cumulative number of alerts delivered to the primary patient group of interest here-to give an indication of the potential clinical impact of the proposed models. The histogram illustrates the cumulative number of unplanned dialysis patients for whom an alert was triggered (imminent kidney failure risk detected), and when it first occurred in their observation period. The reported values should be considered in the context of current clinical performance: 0% detection of unplanned dialysis.

It is important to note that a naïve model predicting 100% kidney failure risk at every timepoint would be perfectly sensitive and deliver alerts to 100% of unplanned dialysis patients. Therefore, fundamental to the formulation of these detection histograms is a counterbalancing measure of precision. A good approach is to take the number of alerts delivered by the models

when operating at specific stepwise precisions [73, 74]. We chose values of 60%, 70%, and 80% based on the performance of our models in our previous study and determined the associated probability cutoff from internal cohort performance.

Sensitivity Analysis

Three sensitivity analyses were performed on the final derived models. Monte Carlo sampling of input features was done to quantify the prediction uncertainty that arises from the natural variability of the laboratory measurements, such as intraday fluctuations in laboratory values. 1,000 samples were generated for each patient visit by applying a noise factor. The noise factor was randomly sampled from a normal distribution with mean zero and standard deviation *v*. We defined *v* as the day-to-day variability. We apply a uniform day-to-day variability of 11.3% to the albumin-to-creatinine ratio, and 6.6% to serum creatinine throughout the Monte Carlo simulations [71]. Second, external validation within independent test cohorts enabled concrete assessment of the validity of the internal cohort results. Finally, SHAP analysis was performed to determine the contribution of trend variables included into the models.

Table B-4 contains the Python libraries used to perform all analyses.

5.2.4 Ethics

The clinical and research activities being reported are consistent with the Principles of the Declaration of Helsinki. All protocols were approved by the Ottawa Health Science Network Research Ethics Board (Protocol ID #20150457-01H). Informed consent requirements were waived due to the retrospective nature of the data. The funding agency played no role in the design, findings, interpretation, or write-up of this study. Our reporting follows guidelines for transparent reporting of artificial intelligence algorithms in medicine (MI-CLAIM Checklist, **Table B-5**).

5.3 Results

5.3.1 Study Populations

The baseline population characteristics and outcomes are summarized in **Table 3-2**. The internal derivation cohort included 1,849 consecutive patients with advanced CKD referred to the Ottawa Hospital (TOH) Multi-Care Kidney Clinic. Two external cohorts of advanced CKD patients were obtained from the Kingston General Hospital (KGH), in Kingston, Ontario, Canada, and the University Health Network (UHN), in Toronto, Ontario, Canada. The external cohorts comprised 1,033 and 323 individuals, respectively. Patients' CKD was the most advanced in the TOH cohort, with mean eGFR (SD) equal to 19 (7) *mL/min/1.73m^2*, compared to 21 (7) *mL/min/1.73m^2* and 23 (8) *mL/min/1.73m^2* for KGH and UHN, respectively. In turn, for the external cohorts this yielded a higher number of patients still being followed at the time of data collection, with 363 (35%) and 182 (56%) patients in KGH and UHN, respectively, vs. 281 (15%) at TOH. The number of patients experiencing kidney failure was also greater in TOH than in the external KGH and UHN cohorts – 1,203 (65%) vs. 471 (46%) and 93 (29%). As was the incidence of unplanned dialysis – 435 (39%) vs. 161 (35%) and 22 (26%).

5.3.2 Internal Performance

Internal performance metrics are tabulated in **Table 5-1**. Performance metrics were studied across several models incorporating different input variables (**Table B6-7**). Supplemental **Tables B8-9** show the performance stratified among subgroups for sex and age quintile. Calibration figures are presented in **Figure 5-1**, panels **A-D**.

The 6-month model achieved an AUC-ROC of 0.878 (95% CI: 0.871 - 0.885), an AUC-PR of 0.689 (95% CI: 0.672 - 0.706), and a Brier score of 0.099 (95% CI: 0.095 – 0.103). 6-month internal calibration is illustrated in **Figure 5-1**, panel **A**.

The 12-month model achieved an AUC-ROC of 0.863 (95% CI: 0.856 - 0.871), an AUC-PR of 0.778 (95% CI: 0.764 - 0.794), and a Brier score of 0.138 (95% CI: 0.133 – 0.142), internally. Internal calibration for the 12-month model is illustrated in **Figure 5-1**, panel **C**.

The inclusion of features measuring time-varying trends in laboratory measurements generally improved performance. The performance increase was most substantial with fewer base laboratory measurements, and negligible with the larger 8-variable laboratory sets. The 6-month model benefited the most from the inclusion of these time-varying trends based upon the relative performance increases within laboratory sets (**Table B-6**).

Performance values shifted with increasing age: discrimination metrics (AUC-PR, AUC-ROC) appeared to worsen, while metrics driven by the prevalence of the majority class (Brier) appeared to improve (**Tables B8-9**).

Table 5-1: Model performance metrics.

	Internal		**External Validation**	
	6-Month Model	**12-Month Model**	**6-Month Model**	**12-Month Model**
Brier Score (95% CI)	0.099 (0.095, 0.103)	0.138 (0.133, 0.142)	0.092 (0.086, 0.097)	0.131 (0.124, 0.137)
AUC-ROC Score (95% CI)	0.878 (0.871, 0.885)	0.863 (0.856, 0.871)	0.872 (0.862, 0.882)	0.868 (0.856, 0.880)
AUC-PR Score (95% CI)	0.689 (0.672, 0.706)	0.778 (0.764, 0.794)	* 0.557 (0.521, 0.589)	0.752 (0.727, 0.779)

Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; AUC-PR, area under the precision-recall curve; CI, confidence interval.
*: Significantly different from internal result, based upon overlapping 95% confidence intervals.

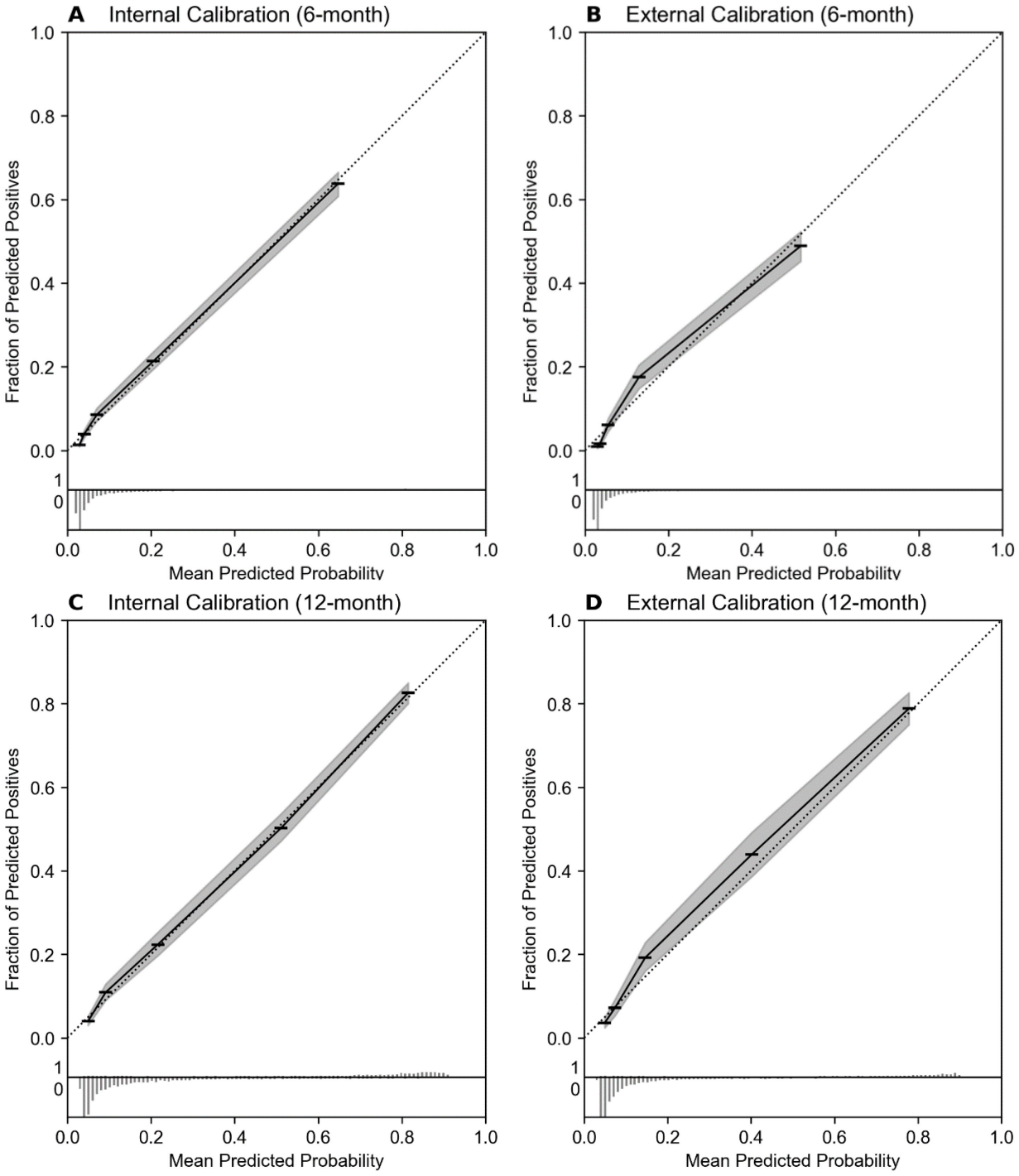


Figure 5-1: Calibration curves for the 6-month model internally (**A**) and upon external validation (**B**), and the 12-month model internally (**C**) and upon external validation (**D**). Min-max-normalized histograms representing the distribution of model predictions is illustrated at the bottom of the figure, with visits falling within 6 or 12 months of a KRT event (1) displayed above, and visits falling outside 6 or 12 months of a KRT event (0) displayed below. Throughout each model training procedure, training was performed on 90% of the training partition, with the remaining 10% being used for model calibration.

5.3.3 External Validation

External validation results for both models are available in **Table 5-1**. Stratified analyses across a sex and age subgroups in the external validation cohorts are presented in supplemental **Tables B10-11**. The 6-month model produced a significantly lower AUC-PR in the external validation set compared to the internal validation set (0.557 [95% CI: 0.521 – 0.589] vs. 0.689 [0.672, 0.706]). Differences between internal and external results were otherwise not significant based upon overlapping 95% confidence intervals. 6-month and 12-month model calibrations in the external cohorts are illustrated in **Figure 5-1**, panel **B** and **Figure 5-1**, panel **D**, respectively.

5.3.4 Impact Measures

Internal impact results for the models are comprehensively presented in supplemental **Tables B12-13** (6-month and 12-month, internal), and **Tables B14-15** (6-month and 12-month, external). The 12-month model delivered timely alerts to 30.6% (95% CI: 26.1% - 35.2%), 48.1% (95% CI: 43.4% - 52.9%), 64.4% (95% CI: 59.9% - 69.0%) of all unplanned dialysis patients, with stepwise precisions of 80%, 70%, and 60%, respectively, at least 3 months prior to their dialysis event (**Figure 5-2**). At the stricter timeframe of 6 months, for the same stepwise precisions, alerts were delivered for 9.7% (95% CI: 7.2% - 12.5%), 20.1% (95% CI: 16.5% - 23.9%), 31.8% (95% CI: 27.4% - 36.1%) of all unplanned dialysis patients respectively. These results persisted upon external validation (no significant difference based upon overlapping 95% CIs). The confusion tables for the 6- and 12-month models at probability cutoffs representing a stepwise precision of 70%, as determined from internal validation, are available in **Table B-16**.

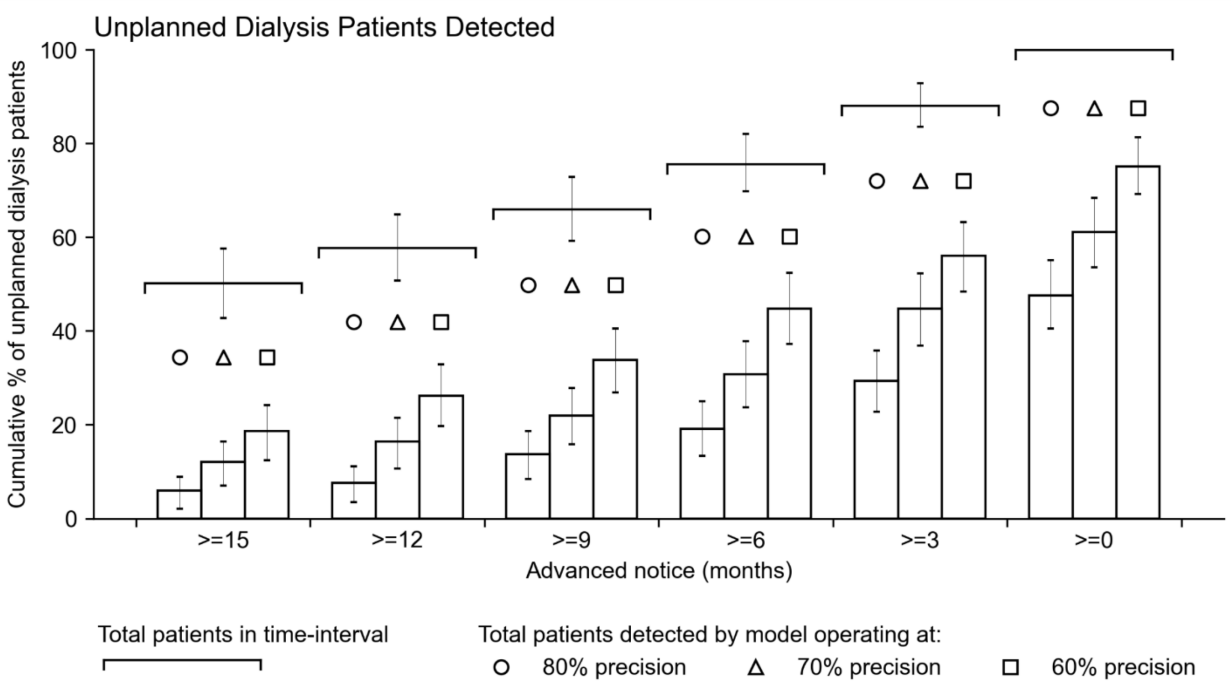


Figure 5-2: For the 12-month model, illustration of the latency between prediction and outcome for those patients that began dialysis in an unplanned manner in the external validation cohorts [74]. The figure only contains unplanned dialysis patients to demonstrate that the model is able to deliver alerts on this challenging subgroup. As such, only positives (dialysis) are represented in this figure. As a measure to counterbalance this, stepwise precisions of 80%, 70%, and 60% are illustrated in each bar cluster to demonstrate how model sensitivity to this subgroup varies. Overhanging bars indicate when unplanned dialysis patients first presented to the clinic. E.g., 50% presented with 15 months or more latency before their outcome. 3-bar clusters underneath plot the cumulative percentage of unplanned dialysis patients predicted with >=X months latency to the outcome (or advanced notice).

5.3.5 Sensitivity analyses

Prediction variability remained under 5% for the majority of cases in both the 6- and 12-month model, as evaluated on the external validation set (**Figure B-3**, panels **A-B**). A correlation between the standard deviation over Monte Carlo prediction iterations was uncovered, in that the closer the prediction was to 0% or 100%, the less the predicted probability was perturbed by noise in the input variables. Further study of this relationship is available in supplemental **Figure B-2** and indicated this uncertainty region was associated with a creatinine range of 350-400 *μmol/L* for the 6-month model, and 300-350 *μmol/L* for the 12-month model. SHAP summary plots are available in supplemental **Figures B4-5**.

5.4 Discussion

In this study, machine learning models were derived to predict kidney failure at short timeframes of 6 and 12 months. The potential for significantly reducing the incidence of unplanned dialysis starts in advanced CKD in three independent advanced CKD cohorts in Ontario, Canada, was evidenced, thereby suggesting the possibility for substantial positive impact on patient lives and provider practice.

The intended goal of these models is to provide decision aid to nephrologists operating in advanced CKD clinics by providing kidney failure alerts in the hope that these alerts may facilitate timely remediation of risk factors for initiating dialysis in an unplanned fashion. The nature of this clinical challenge yields itself to predictive modeling, whereby a patient's immediate or future risk of kidney failure is estimated using statistical algorithms. While there have been numerous kidney failure risk prediction models developed and implemented over the past decade [9-12], the rate of unplanned dialysis starts remains persistently high. These models are decoupled from the problem for two reasons. Firstly, they predict over longer timeframes of 2-5 years, when dialysis preparation should preferably occur 6-9 months prior to dialysis initiation [13]. Second, they are single-timepoint models, derived for more general CKD populations where repeated measures can be more difficult to obtain. A model tailored for prediction in advanced CKD settings would be dynamic and would predict at routine time intervals, similar to the manner and style of patient assessment in specialized CKD clinics. Altogether, this suggests that what may be lacking in advanced CKD practice, is a model that is concretely tied to the described clinical question: must the patient be prepared for dialysis now, or not. I.e., will the patient’s kidneys fail in the next 6-9 months? Our derived models directly tie into this specific clinical question, providing short timeframe (6 and 12 months) risk updates at any timepoint in a patient’s observation period.

Our results indicate that the investigated models could provide significant clinical benefit in this regard. From the example in **Figure 5-2**, it is demonstrated that the model can deliver

timely alerts with high precision to a substantial proportion of unplanned dialysis patients – 20% - 40% – at least 6 months prior to when they started dialysis. The alerts delivered by the models concretely express the need for dialysis at short timeframes so that patients may be prepared for and informed about treatments such as hemodialysis, peritoneal dialysis, kidney transplantation, and conservative management. In general, preparation should occur a few months in advance [99], and studies have shown there is on average no benefit to starting earlier than 6 months [100]. Still, these timeframes will vary from center-to-center meaning that even within our own practice where patients are routinely prepared with 6 months or less time before their dialysis start, cost-benefit analyses will be required to determine optimal treatment pathways at each of the operating points studied. Additionally, given the retrospective nature of this study, we could only evaluate this under idealistic assumptions by simulating when an alert would be delivered for each of the unplanned dialysis patients in our retrospective external cohorts. Therefore, while we hypothesize that these models could help mitigate risks associated with unplanned dialysis initiation by improving patient understanding and facilitating optimal decision-making in the clinical setting [4-6, 8], further analysis is required to determine their true clinical applicability.

We highlight the performance of the models across several identifiable subgroups (**Tables B8-11**). This was of paramount importance to the characterization of our models. A previous landmark machine learning study into the continuous-time prediction of AKI reported greatly improved predictive performance on the Veteran Affairs database [73], a predominantly male population. When one of the models was replicated and applied to a more diverse cohort, significant biases were uncovered [101]. This highlights the importance of deriving models in cohorts that are representative of the target population. All three cohorts studied here, while having their differences, are broadly representative of advanced CKD centers in Ontario, Canada. We show that our models validate throughout, and do not perform unexpectedly in the identifiable groups studied. We attribute the decreasing performance in elevated age quintiles to the rising

competing risk of death in these age groups [47, 102, 103], and a decreasing positive class prevalence. Additionally, each of the three studied cohorts have notable discrepancies in positive class prevalence between each other. This fact represents an important consideration when interpreting our reported model performances across centers.

Our models are not without limitations. Developed in a single-center setting, broader variation in patient demographics and healthcare practice, such as the timing of dialysis initiation, may not be accounted for. External validation results were promising, but the cohorts were comparatively small, and additional validation will be necessary to ensure the models' general applicability. Future research will be crucial to assess the feasibility and cost-effectiveness of integrating these models into regular advanced CKD practice. A critical question will be the cost-benefit analysis of pre-empting more unplanned dialysis initiations versus potentially initiating dialysis preparations earlier than required for some patients. Moreover, while our objective would be to completely eliminate the occurrence of unplanned dialysis among advanced CKD patients through short timeframe kidney failure risk prediction, we realize that this will not be possible. Factors contributing to unplanned dialysis, such as acute kidney injury, lack of nephrology referrals, and some patients' reluctance to consult nephrologists or partake in the kidney failure preparation process are beyond the scope of our model. However, considering the considerable room for improvement (around half of all dialysis initiations), even a modest reduction through improved kidney failure risk prediction could yield significant benefits for both patients and the healthcare system at large.

Chapter 6: Conclusions

Here, the contributions presented in this thesis are summarized. The research and work outlined here represents several marked contributions to the field of kidney failure prediction. Several new insights are offered. And a novel approach is proposed to address this difficult yet crucial clinical problem. The following sections highlight the key findings and areas for future exploration.

6.1 Summary of Contributions

1. One of the principal contributions of this thesis includes a thorough analysis of the extracted clinical dataset followed by the implementation of several transformative procedures. This encompassed a complete cleaning of the dataset, an examination of variable missingness, an imputation analysis, and the uncovering of latent features within the dataset. This portion of the thesis therefore provides a statement of the quality and consistency of the dataset but also unveils concealed insights that enhance the overall comprehension of its intricacies. Such findings may inform the downstream implementation of automated clinical data pipelines in the advanced CKD clinic.
2. On this note, the thesis identified a feature engineering approach (introduced in **Chapter 3** and studied in **Chapter 3** and **Chapter 5**) to explicitly incorporate representations of clinician decision-making processes into the predictive models via a synthesized set of features measuring time-varying trends in laboratory measurements. It was shown that these features augment the predictive accuracy of the models, especially in reduced variable settings. The significant performance boost afforded to a sex, age, and creatinine-only model upon the incorporation of measures of change in creatinine is a particularly exciting finding given the wide availability of these patient data. These techniques may

maximize the accessibility of the proposed clinical tools by providing improved risk assessment in reduced-variable settings.

3. **Chapter 4** introduced four new predictive models designed explicitly for the novel clinical prediction task of short-term kidney failure (6-12 months) among patients with advanced CKD. These models represent an advancement in the field, as there were no existing prediction models tailored to this clinical problem. By concretely tying the models proposed here to the clinical question at hand, they exhibit improved accuracy and suitability for predicting kidney failure within this critical time window. Moreover, the comparison between machine learning approaches and traditional methods demonstrated the superiority of the former. This finding may inform the development of future kidney failure prediction models.
4. **Chapter 5** extensively characterized and validated the best model from **Chapter 4**, ensuring its applicability and potential for mitigating the burden of unplanned dialysis in advanced CKD clinics across Ontario. This chapter therefore represents a comprehensive proposal for new short timeframe kidney failure risk prediction models to address this important clinical problem. Importantly, the results of this retrospective analysis provide a strong indication that upwards of 30% of unplanned dialysis starts could be changed to planned (optimal) dialysis starts. If this potential reduction in suboptimal starts were to carry over into a clinical implementation, the real-world positive impact to both patients and providers would be substantial.

6.2 Limitations

6.2.1 Dataset

Despite the promise of these models, the dataset used in this study poses certain limitations that deserves further attention in future research. One critical limitation is the presence of variable confounding, specifically regarding urine albumin-to-creatinine ratio (uACR). The incorporation of additional relevant clinical predictors such as medication use may help address this confounding issue and improve the models' accuracy and reliability. In the future, deep learning and larger datasets could enable deep learning analyses for phenotyping patients into meaningful drug-response classes. Such information could serve as a clinical aid or as input features into a future model. Finally, the dataset's fluctuating times, compounded by external factors like the COVID-19 pandemic, present challenges to model stability and generalizability. To mitigate this, once again it will be crucial to consider larger datasets and periodic retraining of the models. This ensures that the predictive models remain up-to-date and relevant in dynamic clinical environments.

6.2.2 Outcomes

Another limitation lies in the lack of standardized outcome definitions, in that it is difficult to unify all of the patients that should be flagged for dialysis under a clear and marked endpoint. The endpoints used in this thesis correspond to the endpoints traditionally used in kidney failure prediction research – the time of dialysis initiation [9, 40, 67, 68]. However, the initiation of dialysis and its indicated time for one patient does not necessarily represent the same thing as the initiation of dialysis in another patient. This is true for patients beginning dialysis in a planned manner, and it is especially true for patients beginning dialysis in an unplanned manner. The timing of dialysis is variable, and is contingent upon the clinician’s decision to intervene, the

patient's willingness and availability to start treatment, the clinics scheduling, and potentially other factors such as age and comorbidities [14]. Many patients on a trajectory to initiating dialysis in a planned manner only do so once their eGFR levels drop to 8 or even 5 *mL/min/1.73m*^{2} – well-below the KDIGO classification for kidney failure. It is not ideal to condition a model on this kind of ambiguity. Recently, composite outcomes of dialysis initiation plus eGFR decline are being incorporated into predictive models and clinical trials [48, 104, 105], where the endpoint is taken to have occurred at the first of the two. Such an endpoint is still not ideal. An eGFR decline endpoint can easily change for a patient based on the timing and frequency of follow-up and fluctuations in laboratory measurements. eGFR decline may also be ill-suited for older CKD populations where the incidence of death before dialysis is greatly elevated [104]. Future research efforts will therefore have to be dedicated to enhancing the consistency of the labels on which the model will be conditioned to ensure applicability across different studies and healthcare institutions.

6.2.3 Feature Engineering

The feature engineering approach, while showing promise, has its limitations. Notably, the current implementation involves 0-imputation at initial and potentially second visits. This approach is relatively naive and should be further refined to improve model performance. The boost to a creatinine-only model was significant, but performance tapered off with the inclusion of more laboratory measures. This phenomenon should be further studied. Improving the quality of the repeated measures of uACR may benefit in this regard.

6.3 Recommendations for Future Work

As the predictive models have shown promise and effectiveness in external clinical sites, conducting further external validation on a provincial and international scale will provide additional evidence of their utility and performance across diverse healthcare settings. Additionally, continuous refinement and improvement of the models will be essential to ensure they remain up-to-date and effective. By incorporating feedback from clinicians and healthcare practitioners, iterative model updates can be developed, enhancing their real-world performance.

Integrating predictive models into clinical practice requires a thorough cost-effectiveness analysis. This analysis should encompass various aspects, including the number of unplanned starts of kidney failure prevented, the competing risk of death, and the potential impact the models would have on patient quality of life. By quantifying the economic implications of implementing these models, healthcare decision-makers can make informed choices about their adoption and potential benefits. Furthermore, the cost-effectiveness analysis can help identify specific patient populations that would benefit most from the predictive models, allowing for targeted and efficient implementation in areas with the highest potential impact.

The growing focus on patient engagement in advanced CKD care highlights the need for patient-centered interventions. Developing software systems that leverage machine learning-driven decision support can educate and empower patients to take a more active role in managing their condition. By providing patients with personalized information and treatment recommendations based on the predictive models, these software systems can foster shared decision-making between patients and healthcare providers. This collaborative approach can lead to better treatment adherence, improved patient outcomes, and enhanced patient satisfaction. Moreover, the software systems can serve as educational tools, delivering relevant information to patients about their condition, treatment options, and lifestyle modifications. By enabling patients

to make informed decisions, these systems can contribute to better self-management and overall well-being.

Perhaps, no less significantly, the lessons learnt in this work and methods develop, may apply to any number of other clinical scenarios in which patients are monitored sparsely for disease progression. Thus, this work has the potential for far reaching impact in fields such as oncology, cardiology, and autoimmune diseases.

In conclusion, this thesis has contributed to the field of kidney failure prediction by introducing new predictive models and data modeling approaches for short timeframe prediction of kidney failure in advanced CKD contexts. The identified limitations open avenues for future research and improvement, aiming to address dataset challenges and refine feature engineering methodologies. Overall, this thesis sets the stage for further advancements in predictive modeling and patient care in the field of nephrology. With continued research and collaboration between academia, healthcare professionals, and patients, the potential to make a meaningful impact on CKD management is within reach.

Appendices

Appendix A Chapter 4 Supplemental Material

A.1 Supplemental Methods

Variables

Variables were included in alignment with several validated kidney failure risk prediction models. Using the 4-variable Kidney Failure Risk Equation (KFRE) [10, 49] as a baseline, we defined the 4-variable set to include age, sex, eGFR, and urine albumin-to-creatinine ratio (ACR). The 8-variable set additionally included serum calcium, phosphate, bicarbonate, and albumin to align with the 8-variable KFRE [10, 49]. The 10-variable set additionally included a history of diabetes mellitus and/or hypertension as used by the 6-variable KFRE [10]. The 13-variable set additionally included variables shown in the Veteran Affairs (VA) Model Study [12] to be predictive in an older and advanced CKD population, namely serum potassium, a history of congestive heart failure, and systolic blood pressure. Inclusion of variables in each model is summarized in **Table 3-2**.

To simulate how missing variables would have to be filled in a prospective clinical setting, we forward-filled any missing values using available measurements from prior visits. No transformations were applied to any variables, except mean-centering in the case of Cox regression modeling.

Data from January 2010 through November 2019 were generated from the Siemens Dimension Vista chemistry analyzer. Data from November 2019 through 31 May 2021 were generated using the Roche Cobas. All data were from routine physician-ordered testing in an Accreditation Canada Diagnostics accredited clinical laboratory where minimum performance standards were met or exceeded (e.g., analytical precision, linearity, analytical specificity). Creatinine methods were enzymatic on both platforms and showed excellent agreement (Passing-Bablock regression slope = 0.98 with an intercept of 6.3 µmol/L [0.07 mg/dL]). Urine

albumin, calcium, and phosphate also showed excellent agreement between methods (high R^2 and slope >0.98). Bicarbonate (total carbon dioxide) showed an absolute difference of 1.67 mEq/L (+/- 1.75, 2SD) where the Roche method runs *lower* than the Siemens. Albumin (plasma) showed an absolute difference of 0.4 g/dL (+/- 0.4, 2SD) where the Roche (BCG) method runs *higher* than the Siemens (BCP). eGFR was calculated using the 2021 equation, which does not include a race component [106]. Of the included 1,757 patients, 1,706 had measurements between January 2010 and November 2019. 209 patients had measurements between November 2019 and 31 May 2021. 166 patients had measurements in both periods.

Outcomes

The outcome of interest was kidney failure over short timeframes. We assigned a binary label to each visit. Visits labeled as the positive class were visits that were within 6, 12, 24 months of a kidney failure event ('kidney failure within X-months'), depending on the timeframe. Visits not falling within X-months of kidney failure, or visits for patients who died before kidney failure, were labeled as the negative class (i.e., 'no kidney failure within X-months'). The precise class counts (positive [outcome event within timeframe]:negative [outcome event not within timeframe]) for each of the 6-, 12-, and 24-month datasets were 2562:9231, 4265:7528, and 6330:5463 respectively. For Cox regression fitting, death was a censoring event.

Baseline Cox Regression

As a baseline for this study, we re-implemented the traditional Cox regression methodology employed by the widely-used KFRE and other modern-day kidney failure risk prediction models [9]. The model was fit using the baseline measurements of each patient obtained at their initial clinic visit. Predicted survivor curves were obtained at each visit using the most recent clinical

data. That is, the Cox model was reapplied at each subsequent patient visit, and kidney failure probabilities were extracted at 6-, 12-, 24-month time points.

While traditionally a method for quantifying the effects of variables on hazard, baseline Cox regression (e.g., KFRE) has notably become the most widely used type of kidney failure risk prediction model in modern-day clinical practice. For this study, baseline Cox models were constructed using variable values collected at a patient's initial clinic visit. Though not required, we centered continuous variables using observed sample means for those variables to facilitate comparison of hazard ratios with KFRE (**Table B-5**). As such, the baseline hazard represented a male with no comorbidities, and mean values across all other variables [51]. The cumulative baseline hazard was estimated using the Breslow method from which an estimate of the baseline survivor function ($\check{S}_0(t)$) could be obtained [51, 52]. A predicted survivor curve can then be obtained using the fitted Cox model, representing a longitudinal continuous-time estimation of survival probability for that individual. Individual survival curves were predicted as

$$\check{S}_i(t) = \check{S}_0(t)^{r_i}$$

where r_i represents the individual's predicted risk score as a consequence of the individual's variable values and estimated Cox model parameters. From the predicted individual survivor curve, the compliment of the survival probability is obtained from the time-point of interest to generate a prediction distribution of kidney failure at that timeframe. For example, patient i's predicted survival at 12 months, given their risk score r_i, is computed as

$$\%\ kidney\ failure\ (12\ months) = 100\% \cdot \left(1\ -\ \check{S}_0(12\ month)^{r_i}\right).$$

Time-Varying Cox Regression

Baseline Cox regression models, such as the KFRE, take variables as constant over the study period and do not incorporate more recent follow-up data into the fitting process. To overcome

this limitation, we derived a Cox model with time-varying variables to allow the Cox model to leverage all follow-up measurement data during fitting. Like the baseline Cox model, the most recent clinical data was fed into the model to predict an updated risk score for the patient, from which kidney failure probabilities were derived at each 6-, 12- and 24-month time point. All continuous variables were treated as time-varying in this analysis. Time-varying covariates require knowledge of variable values of all patients still in the risk set (i.e., still being observed at the time-of-event) at each time any patient develops kidney failure. Given that the timing of this event may occur in between visits for other patients, we approximated variable values by considering the most recently obtained measurement as constant over that period (until the following visit). That is, we broke down a patient's observation time into periods marked by their visits and associate their most recent variable values with that entire period [52]. Hazard was then computed on those updated values for patients still in the risk set, as opposed to the baseline measurements as was done with the baseline Cox regression model. It is important to note the difference in interpretation in the estimated parameters in time-varying covariates compared with the approach for baseline covariates. Under baseline covariates the beta parameters can be interpreted as the effect on hazard of unit differences in the covariate, *at time zero*. Under time-varying covariates, the beta parameters assess differences in hazard with respect to covariates at any defined time period [52, 57], making time-varying modeling more conducive to dynamic prediction [72]. To obtain a prediction of kidney failure, we used the formulation given by Altman and De Stavola, [57] for the probability of surviving through an interval $t + h$, conditional on survival to t:

$$\widetilde{P}_i(t, t+h) = \exp\left[-\left\{\widetilde{H}_0(t+h) - \widetilde{H}_0(t)\right\} * r_i\right].$$

$\widetilde{H}_0(t+h)$ represents the baseline cumulative hazard function h months out from the time of the current visit. We take h to be 6, 12, or 24 months. r_i represents the risk score for the patient using their most recently obtained lab measurements.

Random Survival Forest

Random survival forests are an increasingly prominent machine learning approach to survival analysis as a way to incrementally boost performance over traditional Cox methodology while still producing predicted survivor curves [48, 107]. A random survival forest incorporates only baseline visit data (same as the baseline Cox model) into a bagged ensemble of decision trees. Algorithm hyperparameters were specified to 500 decision trees and out-of-bag evaluation. As a measure of regularization, we specified tree growth to a maximum depth of 16, minimum number of samples required by a leaf node to 8, and the minimum number of samples to split an internal node to 4 [108]. Random survival forests employ a log-rank test to determine if a node split produces significantly different survival distributions in the new potential leaf nodes. For a new individual, a predicted survivor curve was generated using a Kaplan-Meier estimate of the samples in the terminal node of that individual's decision path. We then obtained timeframe-specific probabilities from the predicted survivor curves, aggregated over each tree in the ensemble. As with the baseline Cox model, random survival forests were reapplied at each patient follow-up visit, and probabilities were taken from the 6-, 12-, and 24-month time points.

Random Forest Classifier

Frequent and routine follow-up in advanced CKD clinics yields the question of short-timeframe prediction to alternative modeling paradigms, namely classification. Therefore, on each of the variable sets and timeframes we trained an independent random forest classifier using all available visit data to predict a probability of kidney failure 6, 12, 24 months away from each clinic visit. The random forest algorithm is a bagged ensemble method for supervised machine learning [61]. The same hyperparameters were employed as with the random survival forest. Each class was weighted according to its prevalence in the respective dataset to compensate for class imbalances. Random forest classifiers were trained to predict visits 6, 12, or 24 months away from

kidney failure. The trained models produced prediction probability distributions over the relevant timeframe as the arithmetic mean of each individual decision tree prediction.

Statistical Analysis to Compare Model Performance

We performed five-fold cross-validation with stratified cold-patient splits, ensuring each fold contained a class balance representative of the overall data, and all of a patient's visits were contained within a single fold. Given the class imbalance within our 6-month dataset, K=5 was the largest value for K that ensured representative data conditions were maintained in each fold (class balance, unique patients).

While AUC-ROC is more commonly used, AUC-PR better reflects performance in imbalanced data sets, such as is the case here (i.e., more visits without outcome event than with outcome event within the timeframe in question); though no single metric in isolation is sufficient.

However, while AUC-type metrics summarize predictive performance at all possible decision threshold values, many of those threshold values are irrelevant (e.g., using extremely permissive or restrictive thresholds). The PrRe70 captures a model's performance at one point along the PR curve, representing a more concrete statement of potential clinical impact.

A.2 Supplemental Appendix

Supplemental Table A-1. Minimum Information about Clinical Artificial Intelligence Modeling (MI-CLAIM) Checklist.

Study design (Part 1)	**Completed:**	**page number**	**Notes if not completed**
The clinical problem in which the model will be employed is clearly detailed in the paper.	☑	81	
The research question is clearly stated.	☑	81-82	
The characteristics of the cohorts (training and test sets) are detailed in the text.	☑	83-84, 88-89, 11, Tab. 4-1, Supp Tab. A-9	
The cohorts (training and test sets) are shown to be representative of real-world clinical settings.	☑	83-84, 88-89, 11, Tab. 4-1, Supp Tab. A-9	
The state-of-the-art solution used as a baseline for comparison has been identified and detailed.	☑	81-82, 85	
Data and optimization (Parts 2, 3)	**Completed:**	**page number**	**Notes if not completed**
The origin of the data is described and the original format is detailed in the paper.	☑	83	
The independence between training and test sets has been proven in the paper.	☑	86-88	
Transformations of the data before it is applied to the proposed model are described.	☑	Sup. Meth. (Appendix A.1)	
Details on the models that were evaluated and the code developed to select the best model are provided.	☐		Model/variable selection was not performed. Reasonable a priori variable selections and model hyperparameters were selected based on prior relevant literature and machine-learning best-practices.
Is the input data type structured or unstructured?	☑ Structured ☐ Unstructured		
Model performance (Part 4)	**Completed:**	**page number**	**Notes if not completed**
The primary metric selected to evaluate algorithm performance (e.g.: AUC, F-score, etc) including the justification for selection, has been clearly stated.	☑	86-88	

The primary metric selected to evaluate the clinical utility of the model (e.g. PPV, NNT, etc) including the justification for selection, has been clearly stated.	☑	86-88, Sup. Meth. (Appendix A.1)	
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.	☑	89-91. Tab. 4-2	Non-parametric statistics were used by comparing overlapping 95% confidence intervals derived using bootstrapping.
Model Examination (Parts 5)	**Completed:**	**page number**	**Notes if not completed**
Examination Technique 1[a]	☑	101	
Examination Technique 2[a]	☑	101	
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented.	☑	104	
A discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented.	☑	105	
A discussion of the reliability and robustness of the model as the underlying data distribution shifts is included.	☑	106	
Reproducibility (Part 6): choose appropriate tier of transparency			**Notes**
Tier 1: complete sharing of the code	☑		Essential experiment code is shared as a pair of Jupyter notebooks, available on Zenodo at: https://doi.org/10.5281/zenodo.8070935. Certain steps of the experiments are excluded for issues of patient confidentiality or clarity of experiments.
Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation	☐		
Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details	☐		
Tier 4: no sharing	☐		

Abbreviations: PPV, Positive Predictive Value; NNT, Numbers Needed to Treat.

Supplemental Table A-2. Python software libraries used in analyses.

Package	Used For
lifelines (v0.27) [109]	Cox regression modeling.
scikit-learn (v1.02) [63]	Random forest classifier modeling and experiment pipelining.
scikit-survival (v0.17.2) [110]	Random survival forest modeling.
numpy (v1.22.4) [111]	Statistical analyses and data wrangling.
eli5 (v0.13.0)	Permutation importance (Examination 2).
SHAP (v0.41.0) [112]	Computation of Shapley Values and visualization plots (Examination 1).

Supplemental Table A-3. Cross-validation area under the precision-recall curve (AUC-PR) score results of 6, 12, and 24-month models across variable sets (95% confidence intervals) in derivation cohort.

		Timeframe		
		6-Month	12-Month	24-Month
Cox Baseline	4 Variable	0.62 (0.59, 0.64)	0.77 (0.74, 0.79)	0.86 (0.84, 0.89)
	8 Variable	0.62 (0.59, 0.64)	0.76 (0.74, 0.78)	0.86 (0.83, 0.88)
	10 Variable	0.62 (0.60, 0.64)	0.76 (0.75, 0.78)	0.86 (0.84, 0.88)
	13 Variable	0.63 (0.60, 0.65)	0.76 (0.75, 0.78)	0.86 (0.84, 0.88)
Cox Time-Varying	4 Variable	0.70 (0.67, 0.73)	0.79 (0.76, 0.81)	0.85 (0.83, 0.86)
	8 Variable	0.70 (0.67, 0.73)	0.79 (0.75, 0.82)	0.85 (0.83, 0.86)
	10 Variable	0.70 (0.68, 0.73)	0.79 (0.75, 0.82)	0.85 (0.83, 0.86)
	13 Variable	0.71 (0.68, 0.73)	0.79 (0.76, 0.81)	0.85 (0.83, 0.86)
Random Survival Forest	4 Variable	0.68 (0.65, 0.72)	0.79 (0.77, 0.80)	0.87 (0.85, 0.88)
	8 Variable	0.69 (0.65, 0.72)	0.79 (0.77, 0.81)	0.87 (0.85, 0.88)
	10 Variable	0.69 (0.65, 0.72)	0.79 (0.77, 0.81)	0.87 (0.85, 0.89)
	13 Variable	0.69 (0.64, 0.72)	0.79 (0.77, 0.81)	0.87 (0.85, 0.88)
Random Forest Classifier	4 Variable	0.70 (0.67, 0.72)	0.79 (0.77, 0.81)	0.87 (0.86, 0.88)
	8 Variable	0.70 (0.67, 0.72)	0.80 (0.77, 0.82)	0.88 (0.86, 0.89)
	10 Variable	0.70 (0.68, 0.72)	0.80 (0.78, 0.82)	0.88 (0.86, 0.89)
	13 Variable	0.71 (0.68, 0.73)	0.80 (0.78, 0.82)	0.88 (0.86, 0.89)

Supplemental Table A-4. Cross-validation Brier score results of 6, 12, and 24-month models across variable sets (95% confidence intervals) in derivation cohort.

		Timeframe		
		6-Month	12-Month	24-Month
Cox Baseline	4 Variable	0.14 (0.14, 0.14)	0.17 (0.16, 0.17)	0.16 (0.16, 0.17)
	8 Variable	0.14 (0.14, 0.14)	0.17 (0.16, 0.17)	0.16 (0.16, 0.17)
	10 Variable	0.13 (0.12, 0.13)	0.15 (0.15, 0.16)	0.17 (0.16, 0.18)
	13 Variable	0.13 (0.12, 0.13)	0.15 (0.15, 0.16)	0.17 (0.16, 0.18)
Cox Time-Varying	4 Variable	0.11 (0.11, 0.11)	0.16 (0.15, 0.17)	0.22 (0.21, 0.22)
	8 Variable	0.11 (0.10, 0.11)	0.16 (0.15, 0.17)	0.21 (0.21, 0.22)
	10 Variable	0.10 (0.10, 0.11)	0.15 (0.14, 0.16)	0.20 (0.19, 0.21)
	13 Variable	0.10 (0.10, 0.11)	0.15 (0.14, 0.16)	0.19 (0.19, 0.21)
Random Survival Forest	4 Variable	0.11 (0.11, 0.12)	0.14 (0.14, 0.15)	0.16 (0.15, 0.17)
	8 Variable	0.12 (0.12, 0.12)	0.15 (0.14, 0.15)	0.16 (0.15, 0.17)
	10 Variable	0.12 (0.11, 0.12)	0.14 (0.14, 0.15)	0.16 (0.15, 0.17)
	13 Variable	0.12 (0.12, 0.12)	0.15 (0.14, 0.15)	0.16 (0.15, 0.17)
Random Forest Classifier	4 Variable	0.12 (0.12, 0.13)	0.15 (0.14, 0.16)	0.16 (0.15, 0.17)
	8 Variable	0.12 (0.12, 0.12)	0.14 (0.13, 0.15)	0.15 (0.15, 0.16)
	10 Variable	0.12 (0.12, 0.12)	0.14 (0.13, 0.15)	0.15 (0.15, 0.16)
	13 Variable	0.12 (0.12, 0.12)	0.14 (0.13, 0.15)	0.15 (0.15, 0.16)

Supplemental Table A-5. Baseline Cox regression hazard ratios (95% confidence intervals) and p-values indicating significance over HR=1 in derivation cohort.

	Variable Set			
	4 Variable	8 Variable	10 Variable	13 Variable
Age, *per 1 year*	0.977 (0.973, 0.981) P < 0.001	0.978 (0.974, 0.982) P < 0.001	0.976 (0.972, 0.980) P < 0.001	0.974 (0.970, 0.978) P < 0.001
Female Sex	0.776 (0.690, 0.874) P < 0.001	0.798 (0.707, 0.900) P < 0.001	0.802 (0.711, 0.905) P < 0.001	0.794 (0.703, 0.896) P < 0.001
eGFR, *per 1 mL/min/1.73m^2*	0.896 (0.886, 0.906)	0.906 (0.895, 0.917)	0.904 (0.893, 0.916)	0.902 (0.891, 0.914)
Urine Albumin-to-Creatinine Ratio, *per 1 mg/mmol*[a]	1.002 (1.002, 1.002) P < 0.001	1.001 (1.001, 1.002) P < 0.001	1.001 (1.001, 1.002) P < 0.001	1.001 (1.001, 1.002) P < 0.001
Calcium, *per 1 mmol/L*[a]		0.439 (0.292, 0.661) P < 0.001	0.416 (0.276, 0.627) P < 0.001	0.455 (0.300, 0.689) P < 0.001
Phosphate, *per 1 mmol/L*[a]		1.536 (1.256, 1.879) P < 0.001	1.452 (1.183, 1.782) P < 0.001	1.437 (1.169, 1.766) P = 0.001
Bicarbonate, *per 1 mmol/L*[a]		0.987 (0.968, 1.006) P = 0.188	0.985 (0.966, 1.005) P = 0.137	0.980 (0.960, 1.000) P = 0.054
Albumin, *per 1 g/L*[a]		0.98 (0.965, 0.995) P = 0.009	0.981 (0.966, 0.996) P = 0.013	0.976 (0.961, 0.991) P = 0.002
Diabetes Mellitus			1.012 (0.893, 1.146) P = 0.857	0.976 (0.859, 1.108) P = 0.704
Hypertension			1.540 (1.231, 1.927) P < 0.001	1.477 (1.180, 1.850) P = 0.001
Congestive Heart Failure				1.207 (1.041, 1.399) P = 0.013
Potassium, *per 1 mmol/L*[a]				0.999 (0.900, 1.108) P = 0.979
Systolic Blood Pressure, *per 1 mmHg*				1.006 (1.004, 1.009) P < 0.001

Abbreviations: eGFR, estimated glomerular filtration rate.

[a] Laboratory data presented in International System of Units (SI). Conversion to traditional units is as follows: urine albumin-to-creatinine ratio CF 8.85 to convert to mg/g; calcium CF 4.008 to convert to mg/dL; phosphate CF 3.097 to convert to mg/dL; bicarbonate CF 1.0 to convert to mEq/L; albumin CF 0.1 to convert to g/dL; potassium CF 1.0 to convert to mEq/L.

Supplemental Table A-6. Time-varying Cox regression hazard ratios (95% confidence intervals) and p-values indicating significance over HR=1 in derivation cohort.

	Variable Set			
	4 Variable	8 Variable	10 Variable	13 Variable
Age, *per 1 year*	0.985 (0.981, 0.989) P < 0.001	0.985 (0.982, 0.989) P < 0.001	0.984 (0.980, 0.988) P < 0.001	0.982 (0.978, 0.986) P < 0.001
Female Sex	0.706 (0.627, 0.795) P < 0.001	0.720 (0.638, 0.811) P < 0.001	0.725 (0.643, 0.818) P < 0.001	0.728 (0.646, 0.822) P < 0.001
eGFR, *per 1 mL/min/1.73m^2*	0.683 (0.671, 0.696) P < 0.001	0.700 (0.685, 0.714) P < 0.001	0.697 (0.682, 0.712) P < 0.001	0.695 (0.681, 0.710) P < 0.001
Urine Albumin-to-Creatinine Ratio, *per 1 mg/mmol*[a]	1.001 (1.001, 1.001) P < 0.001	1.001 (1.001, 1.001) P < 0.001	1.001 (1.000, 1.001) P < 0.001	1.001 (1.000, 1.001) P < 0.001
Calcium, *per 1 mmol/L*[a]		0.649 (0.463, 0.909) P = 0.012	0.644 (0.459, 0.904) P = 0.011	0.650 (0.462, 0.915) P = 0.014
Phosphate, *per 1 mmol/L*[a]		1.592 (1.342, 1.889) P < 0.001	1.546 (1.301, 1.838) P < 0.001	1.537 (1.291, 1.829) P < 0.001
Bicarbonate, *per 1 mmol/L*[a]		1.016 (0.997, 1.035) P = 0.092	1.012 (0.993, 1.032) P = 0.204	1.008 (0.989, 1.028) P = 0.401
Albumin, *per 1 g/L*[a]		0.975 (0.963, 0.988) P < 0.001	0.975 (0.963, 0.987) P < 0.001	0.974 (0.961, 0.986) P < 0.001
Diabetes Mellitus			1.136 (1.002, 1.288) P = 0.047	1.067 (0.938, 1.213) P = 0.325
Hypertension			1.351 (1.080, 1.692) P = 0.009	1.326 (1.058, 1.662) P = 0.014
Congestive Heart Failure				1.563 (1.345, 1.816) P < 0.001
Potassium, *per 1 mmol/L*[a]				0.950 (0.854, 1.057) P = 0.349
Systolic Blood Pressure, *per 1 mmHg*				1.004 (1.001, 1.007) P = 0.015

Abbreviations: eGFR, estimated glomerular filtration rate.

[a] Laboratory data presented in International System of Units (SI). Conversion to traditional units is as follows: urine albumin-to-creatinine ratio CF 8.85 to convert to mg/g; calcium CF 4.008 to convert to mg/dL; phosphate CF 3.097 to convert to mg/dL; bicarbonate CF 1.0 to convert to mEq/L; albumin CF 0.1 to convert to g/dL; potassium CF 1.0 to convert to mEq/L.

Supplemental Table A-7. Random survival forest variable permutation importance (95% confidence intervals) on Brier score in derivation cohort.

		Variable Set			
		4 Variable	8 Variable	10 Variable	13 Variable
Age	*6 Month* *12 Month* *24 Month*	0.00 (0.00, 0.00) 0.01 (0.01, 0.01) 0.02 (0.01, 0.02)	0.00 (0.00, 0.00) 0.01 (0.01, 0.01) 0.01 (0.01, 0.01)	0.00 (0.00, 0.00) 0.01 (0.00, 0.01) 0.01 (0.01, 0.01)	0.00 (0.00, 0.00) 0.00 (0.00, 0.01) 0.01 (0.01, 0.01)
Female Sex	*6 Month* *12 Month* *24 Month*	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
eGFR	*6 Month* *12 Month* *24 Month*	0.08 (0.08, 0.08) 0.12 (0.12, 0.12) 0.11 (0.10, 0.11)	0.05 (0.05, 0.05) 0.08 (0.08, 0.08) 0.07 (0.07, 0.08)	0.06 (0.06, 0.06) 0.09 (0.08, 0.09) 0.08 (0.08, 0.08)	0.05 (0.05, 0.05) 0.08 (0.07, 0.08) 0.07 (0.07, 0.07)
Urine Albumin-to-Creatinine	*6 Month* *12 Month* *24 Month*	0.02 (0.01, 0.02) 0.03 (0.02, 0.03) 0.04 (0.03, 0.04)	0.01 (0.01, 0.01) 0.02 (0.01, 0.02) 0.02 (0.02, 0.03)	0.01 (0.01, 0.01) 0.02 (0.01, 0.02) 0.03 (0.02, 0.03)	0.01 (0.01, 0.01) 0.01 (0.01, 0.02) 0.02 (0.02, 0.03)
Calcium	*6 Month* *12 Month* *24 Month*		0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
Phosphate	*6 Month* *12 Month* *24 Month*		0.00 (0.00, 0.00) 0.01 (0.00, 0.01) 0.01 (0.01, 0.01)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.01 (0.01, 0.01)	0.00 (0.00, 0.01) 0.00 (0.00, 0.01) 0.01 (0.01, 0.01)

Bicarbonate	*6 Month* *12 Month* *24 Month*		0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (-0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
Albumin	*6 Month* *12 Month* *24 Month*		0.00 (0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
Diabetes Mellitus	*6 Month* *12 Month* *24 Month*			-0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)
Hypertension	*6 Month* *12 Month* *24 Month*			0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)
Congestive Heart Failure	*6 Month* *12 Month* *24 Month*				0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) -0.00 (-0.00, 0.00)
Potassium	*6 Month* *12 Month* *24 Month*				-0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00)
Systolic Blood Pressure	*6 Month* *12 Month* *24 Month*				0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)

Abbreviations: eGFR, estimated glomerular filtration rate.
For each variable and timeframe (6, 12, 24 months), permutation importance is computed as the bootstrap of the sampled mean Brier scores obtained over cross-validation. 95% confidence intervals are expressed within (). Values can be interpreted as the change to the Brier score when the variable is randomly permuted. Larger absolute numbers indicate greater predictive importance, and zero low importance.

Supplemental Table A-8. Random forest classifier variable permutation importance (95% confidence intervals) on Brier score in derivation cohort.

		Variable Set			
		4 Variable	8 Variable	10 Variable	13 Variable
Age	*6 Month* *12 Month* *24 Month*	0.01 (0.01, 0.01) 0.01 (0.01, 0.02) 0.03 (0.02, 0.03)	0.00 (0.00, 0.00) 0.01 (0.01, 0.01) 0.02 (0.02, 0.02)	0.00 (0.00, 0.00) 0.01 (0.01, 0.01) 0.02 (0.02, 0.02)	0.00 (0.00, 0.00) 0.01 (0.01, 0.01) 0.02 (0.01, 0.02)
Female Sex	*6 Month* *12 Month* *24 Month*	0.00 (0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (-0.00, 0.00)
eGFR	*6 Month* *12 Month* *24 Month*	0.12 (0.12, 0.13) 0.14 (0.14, 0.14) 0.12 (0.11, 0.12)	0.09 (0.09, 0.10) 0.11 (0.10, 0.11) 0.09 (0.08, 0.10)	0.09 (0.09, 0.10) 0.11 (0.11, 0.11) 0.09 (0.09, 0.09)	0.09 (0.08, 0.09) 0.10 (0.10, 0.11) 0.08 (0.08, 0.09)
Urine Albumin-to-Creatinine	*6 Month* *12 Month* *24 Month*	0.01 (0.01, 0.02) 0.03 (0.03, 0.03) 0.04 (0.03, 0.04)	0.01 (0.0, 0.01) 0.02 (0.02, 0.02) 0.03 (0.02, 0.03)	0.01 (0.0, 0.01) 0.02 (0.02, 0.02) 0.03 (0.02, 0.03)	0.01 (0.01, 0.01) 0.02 (0.02, 0.02) 0.03 (0.02, 0.03)
Calcium	*6 Month* *12 Month* *24 Month*		0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)
Phosphate	*6 Month* *12 Month* *24 Month*		0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) 0.00 (0.00, 0.01) 0.00 (0.00, 0.01)

Bicarbonate	*6 Month* *12 Month* *24 Month*		-0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)
Albumin	*6 Month* *12 Month* *24 Month*		-0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
Diabetes Mellitus	*6 Month* *12 Month* *24 Month*			-0.00 (-0.00, -0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)	-0.00 (-0.00, -0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
Hypertension	*6 Month* *12 Month* *24 Month*			0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)
Congestive Heart Failure	*6 Month* *12 Month* *24 Month*				0.00 (0.00, 0.00) 0.00 (0.00, 0.00) 0.00 (-0.00, 0.00)
Potassium	*6 Month* *12 Month* *24 Month*				0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)
Systolic Blood Pressure	*6 Month* *12 Month* *24 Month*				-0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)

Abbreviations: eGFR, estimated glomerular filtration rate.
For each variable and timeframe (6, 12, 24 months), permutation importance is computed as the bootstrap of the sampled mean Brier scores obtained over cross-validation. 95% confidence intervals are expressed within (). Values can be interpreted as the change to the Brier score when the variable is randomly permuted. Larger absolute numbers indicate greater predictive importance, and zero low importance.

Supplemental Table A-9. Baseline characteristics of external validation cohort.

	Kingston General Hospital (N = 493; 2016-2023)	University Health Network, Toronto (N = 209; 2015-2023)
Variable	**Summary Statistics**	**Summary Statistics**
Demographics		
Age, *Years, Mean (SD)*	69 (14)	66 (18)
Male Sex, *N (%)*	300 (61)	120 (57)
Laboratory Data[a]		
Creatinine, *mg/dL, Mean (SD)*	3.37 (1.21)	2.91 (1.09)
eGFR, *mL/min/1.73m^2, Mean (SD)*	19 (7)	23 (8)
Urine Albumin-to-Creatinine Ratio, *mg/g, Median (IQR)*	1336 (389, 3132)	841 (292, 2292)
Calcium, *mg/dL, Mean (SD)*	9.10 (0.68)	9.18 (0.60)
Phosphate, *mg/dL, Mean (SD)*	4.24 (0.99)	4.09 (0.90)
Bicarbonate, *mEq/L, Mean (SD)*	23 (4)	23 (4)
Albumin, *g/dL, Mean (SD)*	3.4 (0.6)	3.9 (0.5)
Comorbidities, N (%)		
Diabetes Mellitus	323 (66)	119 (57)
Hypertension	434 (88)	200 (96)
Congestive Heart Failure	93 (19)	32 (15)
Outcomes, N (%)		
Kidney Failure	361 (73)	91 (44)
Death Before Kidney Failure	132 (27)	44 (21)
Still Followed	0 (0)	74 (35)

Abbreviations: eGFR, estimated glomerular filtration rate; IQR, interquartile range; N, number; SD, standard deviation.

[a] Laboratory data presented in traditional units. Conversion to International System of Units (SI) is as follows: creatinine conversion factor (CF) 88.42 to convert to µmol/L; urine albumin-to-creatinine ratio CF 0.113 to convert to mg/mmol; calcium CF 0.2495 to convert to mmol/L; phosphate CF 0.3229 to convert to mmol/L; bicarbonate CF 1.0 to convert to mmol/L; albumin CF 10.0 to convert to g/L.

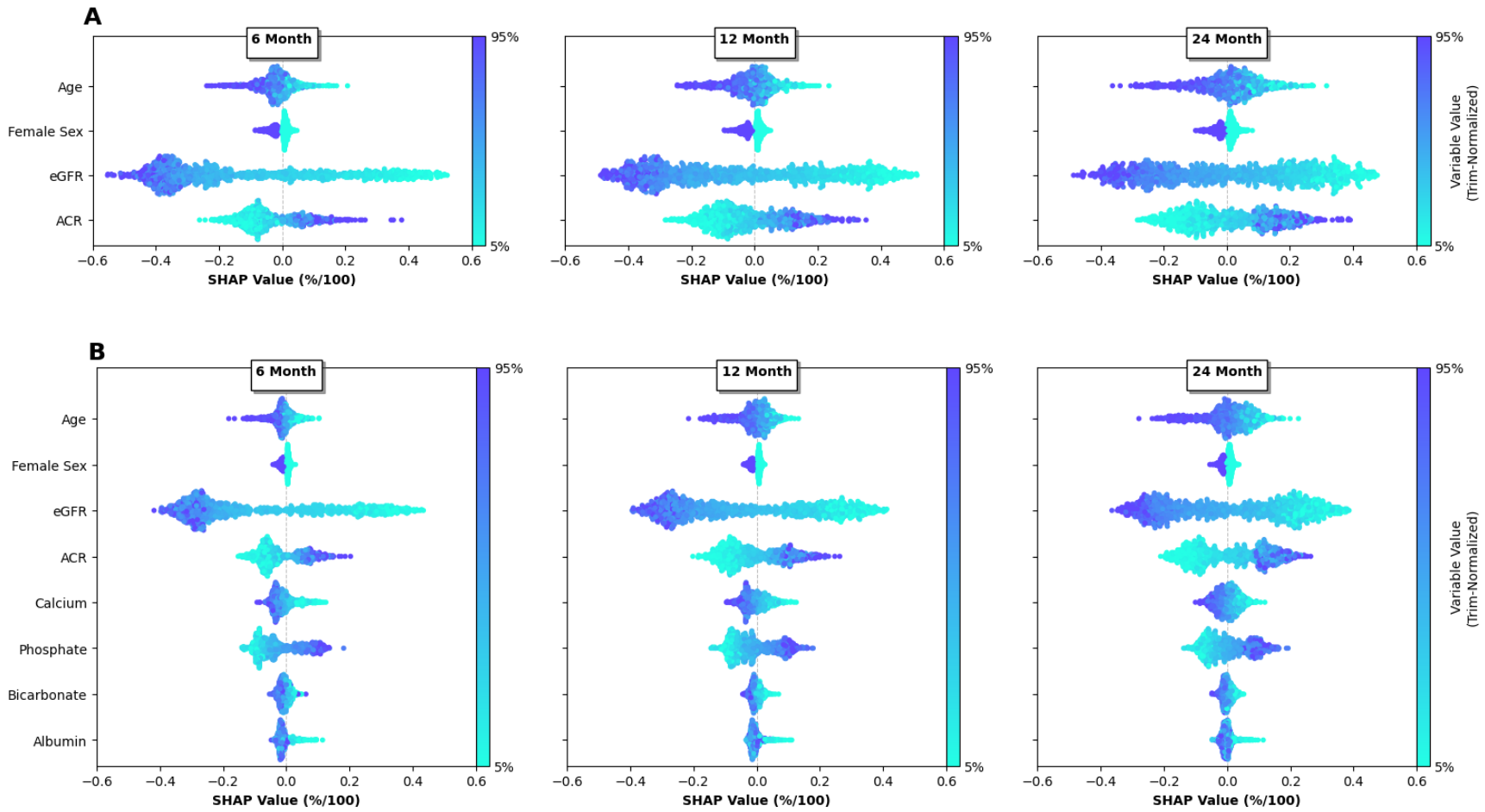
Supplemental Table A-10. External testing area under the precision-recall curve (AUC-PR) score results of selected 6, 12, and 24-month models (95% confidence intervals) in validation cohort.

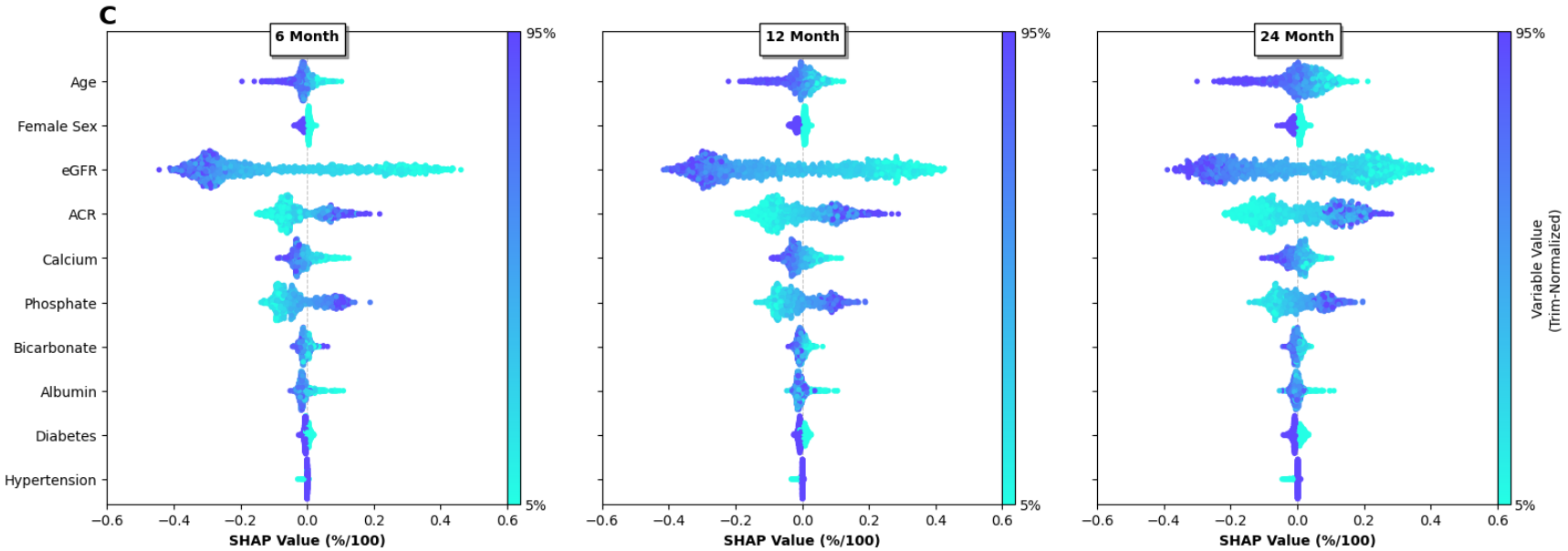
		Timeframe		
		6-Month	12-Month	24-Month
Cox Baseline	4 Variable	0.70 (0.66, 0.73)	0.83 (0.80, 0.86)	0.90 (0.88, 0.92)
Cox Time-Varying	8 Variable	0.71 (0.67, 0.75)	0.81 (0.78, 0.84)	0.86 (0.84, 0.89)
Random Survival Forest	8 Variable	0.71 (0.68, 0.74)	0.82 (0.79, 0.85)	0.88 (0.86, 0.90)
Random Forest Classifier	8 Variable	0.72 (0.68, 0.75)	0.82 (0.79, 0.85)	0.89 (0.86, 0.91)

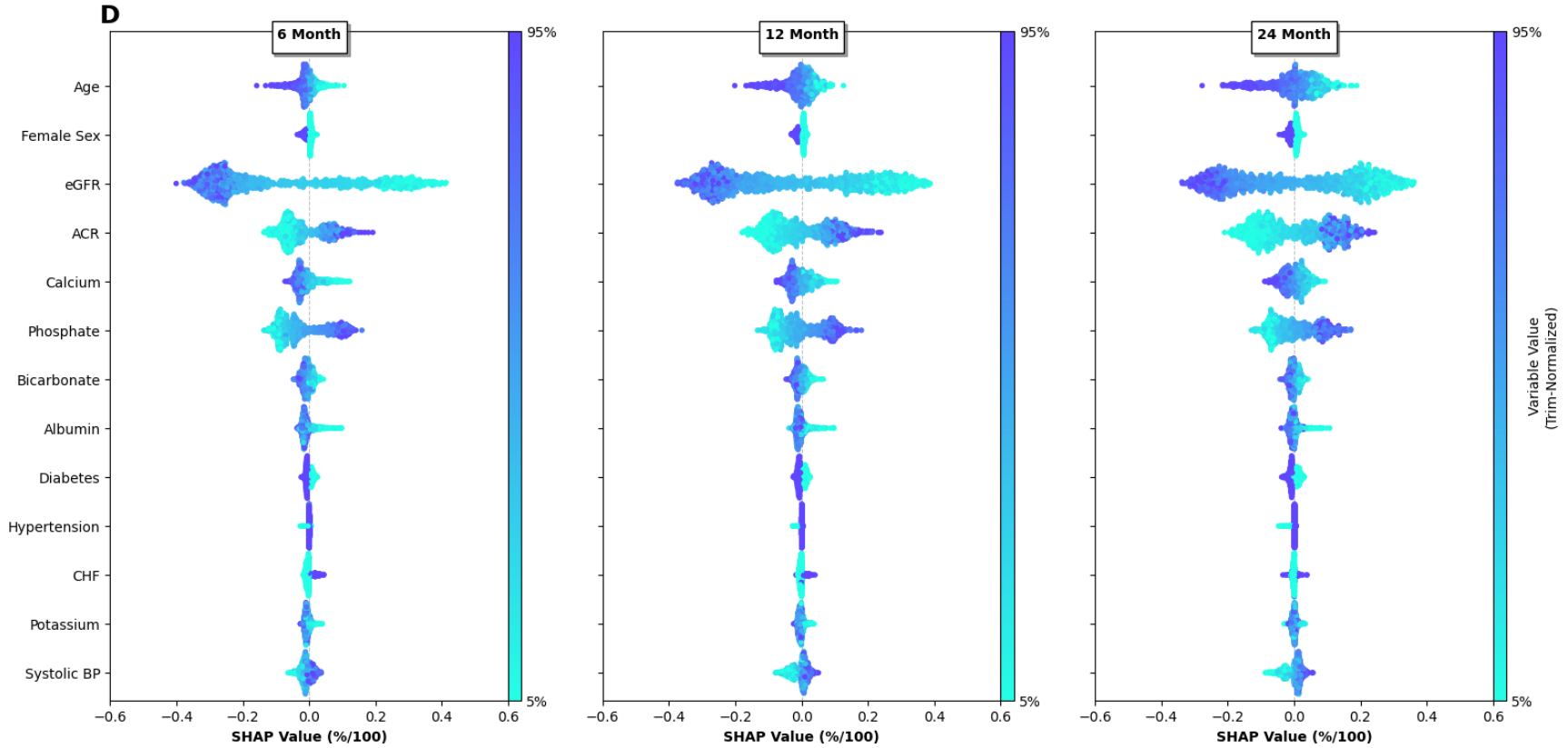
Supplemental Table A-11. External testing Brier score results of selected 6, 12, and 24-month models (95% confidence intervals) in validation cohort.

		Timeframe		
		6-Month	12-Month	24-Month
Cox Baseline	4 Variable	0.17 (0.16, 0.18)	0.18 (0.17, 0.20)	0.16 (0.15, 0.17)
Cox Time-Varying	8 Variable	0.16 (0.15, 0.17)	0.22 (0.20, 0.23)	0.26 (0.24, 0.28)
Random Survival Forest	8 Variable	0.15 (0.14, 0.16)	0.17 (0.16, 0.18)	0.17 (0.15, 0.18)
Random Forest Classifier	8 Variable	0.13 (0.12, 0.14)	0.15 (0.14, 0.17)	0.17 (0.15, 0.18)

Supplemental Figure A-1. SHAP analysis on random forest classifier output (%/100).

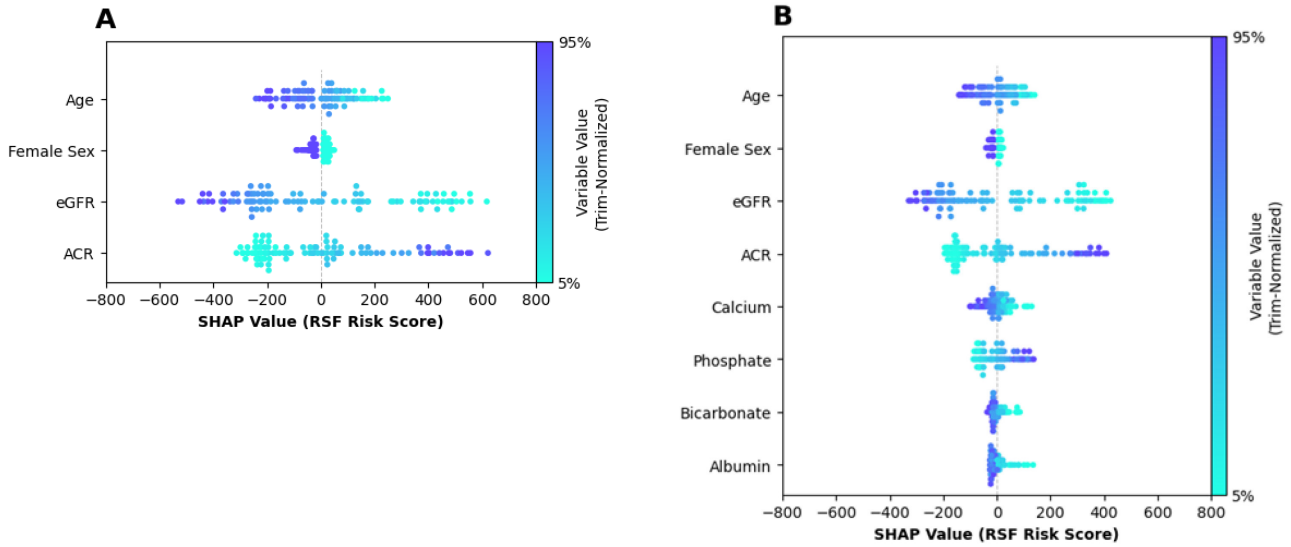


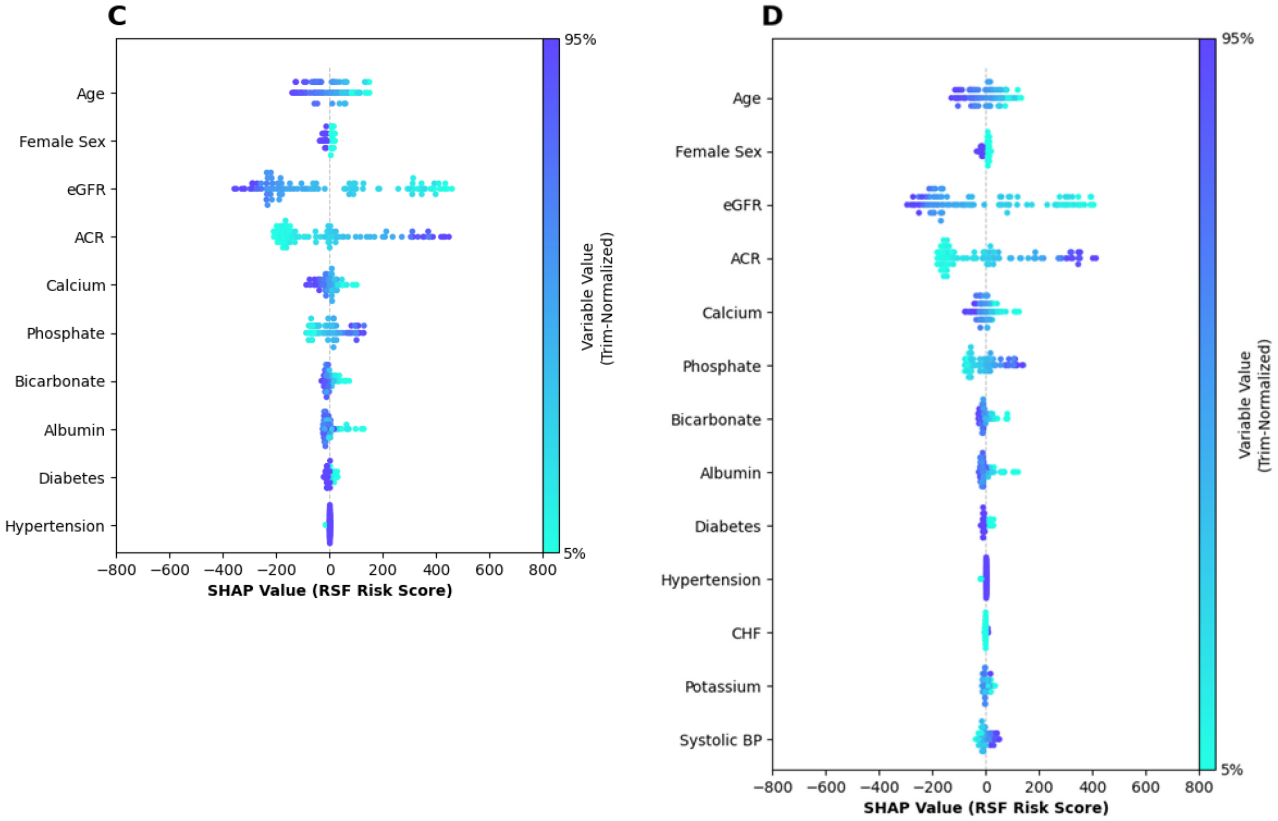

C
6 Month
12 Month
24 Month
Age
Female Sex
eGFR
ACR
Calcium
Phosphate
Bicarbonate
Albumin
Diabetes
Hypertension
95%
5%
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
SHAP Value (%/100)
Variable Value
(Trim-Normalized)



Abbreviations: eGFR, estimated glomerular filtration rate; ACR, urine albumin to creatinine ratio; CHF, Congestive Heart Failure; Systolic Blood Pressure.
SHAP swarmplots for the (A) 4-variable, (B) 8-variable, (C) 10-variable, (D) 13-variable random forest classifier models at each timeframe (6, 12, and 24 months). We explain the same random sample (n=1000) for each model with respect to included variables. Variables are ordered based on inclusion in each variable set. Coded into color is the raw variable value for each example in the random sample using a trimmed normalization (5-95%) to avoid skewing from outliers. Encoded into the x-axis is the SHAP value, representing the mean marginal contribution of that variable value to the model output. In Figures A-D, a strong predictor will be greater-spread in the x-axis and have a uniformly diverging color distribution.

Supplemental Figure A-2. SHAP analysis on random survival forest output (risk score).





Abbreviations: eGFR, estimated glomerular filtration rate; ACR, urine albumin to creatinine ratio; CHF, Congestive Heart Failure; Systolic Blood Pressure.
SHAP swarmplots for the (A) 4-variable, (B) 8-variable, (C) 10-variable, (D) 13-variable random survival forest models using predicted risk score. We explain the same random sample (n=100) for each model with respect to included variables. Variables are ordered based on inclusion in each variable set. Coded into color is the raw variable value for each example in the random sample using a trimmed normalization (5-95%) to avoid skewing from outliers. Encoded into the x-axis is the SHAP value, representing the mean marginal contribution of that variable value to the model output. In Figures A-D, a strong predictor will be greater-spread in the x-axis and have a uniformly diverging color distribution.

Appendix B Chapter 5 Supplemental Material

B.1 Supplemental Appendix

Supplemental Table B-1. Description of trend and change in laboratory measurements.

ID	Description of Property
max	Maximum over entire history.
Δmax	Difference from max.
Δtmax	Difference from max scaled by time.
min	Minimum over entire history.
Δmin	Difference from min.
Δtmin	Difference from min scaled by time.
Δb	Difference from baseline.
Δtb	Difference from baseline scaled by time.
mean3	Moving 3-visit average.
mean	Average over entire history.
Δtp	First difference between visits scaled by time.
std(Δtp)	Standard deviation Δtp over entire history.
mean(Δtp)	Average of Δtp over entire history.
a	Instantaneous acceleration.
mean(a)	Average of a over entire history.

Supplemental Table B-2. Urine albumin-to-creatinine ratio missingness tests.

<table>
<tr><th>Test</th><th>Test Description</th><th>Result</th></tr>
<tr><td>Independence between the missingness of ACR and other features.</td><td>Create a contingency table of ACR missingness against missingness in other features. Perform X^2 test with two degrees of freedom.</td><td>
<table>
<tr><td></td><td>Other Observed</td><td>Other Missing</td></tr>
<tr><td>ACR Missing</td><td>266</td><td>51</td></tr>
<tr><td>ACR Observed</td><td>1620</td><td>64</td></tr>
</table>
p = <0.001

We observed an association between ACR missingness and the missingness of one or more other features.
</td></tr>
<tr><td>Missing at Random (MAR)</td><td>Fit logistic regression model between missingness indicator for ACR and the observed values of other features.</td><td>Female Sex: p = 0.001
Age: p = 0.20
Creatinine: p < 0.001
Year seen: p < 0.001
Season (quarter): p = 0.87

There is an association between ACR missingness and female sex, elevated baseline creatinine, and the year the patient was first seen.</td></tr>
</table>

Supplemental Table B-3. Random forest classifier hyperparameter space.

Hyperparameter	Search Space
n_estimators	[250‡, 500†]
min_samples_leaf	[4, 8†‡, 12]
min_samples_split	[2†, 5, 10‡]
max_depth	[None, 8‡, 16, 24†]
max_features	[1‡, 2, "sqrt"†]

† Optimal for 6-month model.
‡ Optimal for 12-month model.

Supplemental Table B-4. Python packages used.

Task	Package(s)
Pipelining, data manipulation, statistical analysis	pandas, numpy, scikit-learn, scipy
Random forest build	scikit-learn
Figures	matplotlib

Supplemental Table B-5. Minimum Information about Clinical Artificial Intelligence Modeling (MI-CLAIM) Checklist.

Study design (Part 1)	**Completed: page number**		**Notes if not completed**
The clinical problem in which the model will be employed is clearly detailed in the paper.	☑	96	
The research question is clearly stated.	☑	96	
The characteristics of the cohorts (training and test sets) are detailed in the text.	☑	Table 3-2	
The cohorts (training and test sets) are shown to be representative of real-world clinical settings.	☑	Table 3-2	
The state-of-the-art solution used as a baseline for comparison has been identified and detailed.	☑	100	
Data and optimization (Parts 2, 3)	**Completed: page number**		**Notes if not completed**
The origin of the data is described and the original format is detailed in the paper.	☑	98	
Transformations of the data before it is applied to the proposed model are described.	☑	98	
The independence between training and test sets has been proven in the paper.	☑	99	
Details on the models that were evaluated and the code developed to select the best model are provided.	☑	99	
Is the input data type structured or unstructured?	☑ Structured ☐ Unstructured		
Model performance (Part 4)	**Completed: page number**		**Notes if not completed**
The primary metric selected to evaluate algorithm performance (eg: AUC, F-score, etc) including the justification for selection, has been clearly stated.	☑	100	
The primary metric selected to evaluate the clinical utility of the model (eg PPV, NNT, etc) including the justification for selection, has been clearly stated.	☑	100	
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.	☑	Table 5-1	

Model Examination (Parts 5)	Completed: page number		Notes if not completed
Examination Technique 1[a]	☑	101	
Examination Technique 2[a]	☑	101	
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented.	☑	104	
A discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented.	☑	105	
A discussion of the reliability and robustness of the model as the underlying data distribution shifts is included.	☑	106	
Reproducibility (Part 6): choose appropriate tier of transparency			**Notes**
Tier 1: complete sharing of the code	☑		https://github.com/OttawaNMMI/AcuteOnCKD.
Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation	☐		
Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details	☐		
Tier 4: no sharing	☐		

Supplemental Table B-6. Internal performance metrics by variable set (6-month model).

	Brier Score (95% CI)	AUC-ROC (95% CI)	AUC-PR (95% CI)
Clin Chem (8V)	0.117 (0.113, 0.121)	0.878 (0.871, 0.886)	0.679 (0.660, 0.698)
3V	0.132 (0.127, 0.137)	0.855 (0.847, 0.864)	0.630 (0.609, 0.652)
3V + Trends	0.119 (0.114, 0.123)	0.871 (0.863, 0.878)	0.678 (0.661, 0.696)
4V	0.120 (0.116, 0.125)	0.872 (0.864, 0.880)	0.673 (0.654, 0.691)
4V + Trends	0.113 (0.109, 0.117)	0.880 (0.872, 0.886)	0.696 (0.679, 0.713)
8V	0.117 (0.113, 0.121)	0.878 (0.871, 0.885)	0.673 (0.654, 0.692)
8V + Trends	0.110 (0.107, 0.114)	0.884 (0.877, 0.891)	0.691 (0.674, 0.709)

Abbreviations: Clin Chem (8V), random forest classifier developed in the previous study incorporating 8 variables, including estimated glomerular filtration rate (eGFR), age, sex, urine albumin-to-creatinine ratio, phosphate, bicarbonate, calcium, and albumin; 3V, variable set with creatinine, age, sex; 4V, variable set with creatinine, age, sex, and urine albumin-to-creatinine ratio; 8V, same variable set as Clin Chem (8V) except with eGFR substituted for creatinine; Trends, denotes that features measuring change in the base variable set were synthesized and included into the model.

Supplemental Table B-7. Internal performance metrics by variable set (12-month model).

	Brier Score (95% CI)	AUC-ROC (95% CI)	AUC-PR (95% CI)
Clin Chem (8V)	0.143 (0.138, 0.148)	0.863 (0.855, 0.872)	0.781 (0.768, 0.795)
3V	0.159 (0.154, 0.164)	0.835 (0.825, 0.845)	0.731 (0.713, 0.75)
3V + Trends	0.15 (0.145, 0.155)	0.85 (0.842, 0.858)	0.762 (0.746, 0.777)
4V	0.146 (0.141, 0.151)	0.857 (0.849, 0.866)	0.77 (0.755, 0.785)
4V + Trends	0.143 (0.138, 0.147)	0.864 (0.856, 0.872)	0.784 (0.77, 0.798)
8V	0.143 (0.138, 0.148)	0.863 (0.855, 0.872)	0.78 (0.767, 0.794)
8V + Trends	0.141 (0.136, 0.146)	0.867 (0.859, 0.876)	0.782 (0.768, 0.796)

Abbreviations: Clin Chem (8V), random forest classifier developed in the previous study incorporating 8 variables, including estimated glomerular filtration rate (eGFR), age, sex, urine albumin-to-creatinine ratio, phosphate, bicarbonate, calcium, and albumin; 3V, variable set with creatinine, age, sex; 4V, variable set with creatinine, age, sex, and urine albumin-to-creatinine ratio; 8V, same variable set as Clin Chem (8V) except with eGFR substituted for creatinine; Trends, denotes that features measuring change in the base variable set were synthesized and included into the model.

Supplemental Table B-8. Internal performance metrics stratified by group (6-month model).

Group	Brier Score (95% CI)	AUC-ROC (95% CI)	AUC-PR (95% CI)
Sex			
Male (N=1159)	0.101 (0.096, 0.105)	0.876 (0.868, 0.885)	0.697 (0.676, 0.719)
Female (N=690)	0.096 (0.090, 0.102)	0.887 (0.876, 0.898)	0.694 (0.661, 0.726)
Age Quintiles			
18-55 (Q1; N=370)	0.109 (0.102, 0.118)	0.889 (0.875, 0.902)	0.769 (0.739, 0.799)
55-64 (Q2; N=370)	0.105 (0.097, 0.112)	0.883 (0.871, 0.895)	0.723 (0.691, 0.752)
64-72 (Q3; N=369)	0.100 (0.092, 0.109)	0.874 (0.857, 0.890)	0.679 (0.638, 0.724)
72-79 (Q4; N=370)	0.094 (0.085, 0.103)	0.859 (0.839, 0.878)	0.593 (0.544, 0.643)
79-98 (Q5; N=370)	0.084 (0.075, 0.093)	0.864 (0.842, 0.886)	0.546 (0.477, 0.612)

Abbreviations: Q1-5, quintiles 1 through 5; N, number; AUC-ROC, area under the receiver operating characteristic curve; AUC-PR, area under the precision recall curve; CI, confidence interval.

Supplemental Table B-9. Internal performance metrics stratified by group (12-month model).

Group	Brier Score (95% CI)	AUC-ROC Score (95% CI)	AUC-PR Score (95% CI)
Sex			
Male (N=1145)	0.138 (0.132, 0.143)	0.866 (0.857, 0.876)	0.787 (0.769, 0.807)
Female (N=682)	0.137 (0.130, 0.145)	0.869 (0.856, 0.882)	0.780 (0.756, 0.805)
Age Quintiles			
18-55 (Q1; N=366)	0.143 (0.134, 0.152)	0.871 (0.855, 0.886)	0.846 (0.824, 0.868)
55-64 (Q2; N=365)	0.139 (0.130, 0.148)	0.870 (0.856, 0.885)	0.820 (0.797, 0.845)
64-72 (Q3; N=365)	0.142 (0.131, 0.152)	0.856 (0.839, 0.876)	0.775 (0.739, 0.809)
72-79 (Q4; N=365)	0.135 (0.124, 0.145)	0.852 (0.834, 0.870)	0.693 (0.647, 0.745)
79-98 (Q5; N=366)	0.127 (0.116, 0.139)	0.842 (0.818, 0.864)	0.613 (0.552, 0.677)

Abbreviations: Q1-5, quintiles 1 through 5; N, number; AUC-ROC, area under the receiver operating characteristic curve; AUC-PR, area under the precision recall curve; CI, confidence interval.

Supplemental Table B-10. External performance metrics stratified by group (6-month model).

Group	Brier (95% CI)	AUC-ROC (95% CI)	AUC-PR (95% CI)
Sex			
Male (N=826)	0.085 (0.079, 0.091)	0.887 (0.874, 0.898)	0.576 (0.535, 0.613)
Female (N=530)	0.103 (0.093, 0.112)	0.858 (0.841, 0.876)	0.539 (0.481, 0.597)
Age Quintiles			
19-57 (Q1; N=271)	0.115 (0.104, 0.129)	0.882 (0.861, 0.902)	0.698 (0.644, 0.750)
57-68 (Q2; N=271)	0.111 (0.100, 0.123)	0.859 (0.836, 0.881)	0.532 (0.462, 0.607)
68-75 (Q3; N=271)	0.087 (0.075, 0.098)	0.857 (0.831, 0.881)	0.482 (0.400, 0.565)
75-83 (Q4; N=266)	0.078 (0.066, 0.089)	0.861 (0.835, 0.888)	0.444 (0.362, 0.524)
83-100 (Q5; N=277)	0.061 (0.050, 0.071)	0.854 (0.814, 0.889)	0.356 (0.246, 0.466)
Cohorts			
KGH, Kingston General Hospital (N=1033)	0.098 (0.091, 0.104)	0.860 (0.847, 0.872)	0.529 (0.491, 0.568)
UHN, University Health Network (N=323)	0.070 (0.059, 0.080)	0.919 (0.902, 0.936)	0.675 (0.610, 0.730)

Abbreviations: Q1-5, quintiles 1 through 5; N, number; KGH, Kingston General Hospital; UHN, University Health Network Toronto; AUC-ROC, area under the receiver operating characteristic curve; AUC-PR, area under the precision recall curve; CI, confidence interval.

Supplemental Table B-11. External performance metrics stratified by group (12-month model).

Group	Brier Score (95% CI)	AUC-ROC Score (95% CI)	AUC-PR Score (95% CI)
Sex			
Male (N=783)	0.126 (0.117, 0.135)	0.874 (0.858, 0.888)	0.751 (0.721, 0.780)
Female (N=505)	0.138 (0.126, 0.149)	0.872 (0.853, 0.890)	0.779 (0.733, 0.831)
Age Quintiles			
19-57 (Q1; N=258)	0.147 (0.133, 0.161)	0.87 (0.848, 0.892)	0.829 (0.793, 0.862)
57-68 (Q2; N=257)	0.143 (0.126, 0.159)	0.866 (0.840, 0.895)	0.759 (0.692, 0.829)
68-75 (Q3; N=258)	0.134 (0.118, 0.150)	0.867 (0.840, 0.896)	0.797 (0.750, 0.837)
75-82 (Q4; N=257)	0.123 (0.109, 0.139)	0.845 (0.815, 0.875)	0.686 (0.622, 0.747)
82-100 (Q5; N=258)	0.098 (0.084, 0.113)	0.86 (0.824, 0.893)	0.480 (0.361, 0.594)
Cohorts			
KGH, Kingston General Hospital (N=982)	0.138 (0.130, 0.146)	0.856 (0.843, 0.870)	0.744 (0.715, 0.775)
UHN, University Health Network (N=306)	0.098 (0.085, 0.110)	0.915 (0.894, 0.937)	0.797 (0.739, 0.846)

Abbreviations: Q1-5, quintiles 1 through 5; N, number; KGH, Kingston General Hospital; UHN, University Health Network Toronto; AUC-ROC, area under the receiver operating characteristic curve; AUC-PR, area under the precision recall curve; CI, confidence interval.

Supplemental Table B-12. Cumulative % of unplanned dialysis patients for which an alert was triggered by time period (6-month model; internal evaluation).

	Time to Dialysis Event (months)					
	>=15	***>=12***	***>=9***	***>=6***	***>=3***	***>=0***
Total UD Patients (95% CI)	46.4 (41.6, 51.3)	56.3 (51.7, 61.1)	64.1 (59.2, 68.6)	74.2 (70.0, 78.2)	85.5 (82.1, 88.7)	100.0 (100.0, 100.0)
Detected with 60% Precision (95% CI)	4.6 (2.6, 6.8)	8.8 (6.1, 11.6)	11.7 (8.6, 14.9)	18.7 (14.9, 22.4)	31.8 (27.4, 36.1)	57.5 (53.0, 62.2)
Detected with 70% Precision (95% CI)	1.6 (0.6, 3.0)	4.1 (2.4, 6.1)	5.8 (3.8, 8.1)	9.9 (7.2, 12.9)	20.1 (16.5, 23.9)	45.8 (41.1, 50.5)
Detected with 80% Precision (95% CI)	0.5 (0.0, 1.2)	1.1 (0.2, 2.2)	2.7 (1.2, 4.3)	4.6 (2.7, 6.7)	9.7 (7.2, 12.5)	29.7 (25.5, 34.1)

Abbreviations: UD, unplanned dialysis; CI, confidence interval.

Supplemental Table B-13. Cumulative % of unplanned dialysis patients for which an alert was triggered by time period (12-month model; internal evaluation).

	Time to Dialysis Event (months)					
	>=15	***>=12***	***>=9***	***>=6***	***>=3***	***>=0***
Total UD Patients (95% CI)	46.4 (41.8, 51.1)	56.3 (51.9, 60.6)	64.1 (59.4, 68.5)	74.2 (70.2, 78.2)	85.5 (82.0, 88.6)	100.0 (100.0, 100.0)
Detected with 60% Precision (95% CI)	22.5 (18.8, 26.4)	29.6 (25.4, 34.0)	39.3 (34.7, 44.0)	50.8 (46.1, 55.4)	64.4 (59.9, 69.0)	85.8 (82.5, 89.3)
Detected with 70% Precision (95% CI)	10.8 (7.9, 14.0)	16.7 (13.2, 20.4)	23.2 (19.0, 27.5)	35.4 (30.8, 40.2)	48.1 (43.4, 52.9)	72.9 (68.7, 77.2)
Detected with 80% Precision (95% CI)	3.7 (1.9, 5.6)	8.1 (5.6, 10.6)	12.4 (9.4, 15.7)	18.9 (15.1, 22.7)	30.6 (26.1, 35.2)	53.6 (48.8, 58.7)

Abbreviations: UD, unplanned dialysis; CI, confidence interval.

Supplemental Table B-14. Cumulative % of unplanned dialysis patients for which an alert was triggered by time period (6-month model; external evaluation).

	Time to Dialysis Event (months)					
	>=15	***>=12***	***>=9***	***>=6***	***>=3***	***>=0***
Total UD Patients (95% CI)	50.0 (43.1, 57.4)	57.6 (50.3, 64.9)	65.7 (59.0, 72.4)	75.5 (69.3, 81.9)	88.0 (83.4, 92.3)	100.0 (100.0, 100.0)
Detected with 60% Precision (95% CI)	4.4 (1.7, 7.4)	6.0 (2.8, 9.4)	11.9 (7.4, 16.6)	15.1 (9.9, 20.6)	28.1 (21.6, 34.5)	46.7 (39.3, 54.3)
Detected with 70% Precision (95% CI)	1.1 (0.0, 2.9)	3.2 (1.1, 5.9)	5.4 (2.5, 8.7)	9.2 (5.3, 13.5)	18.9 (13.0, 24.9)	37.5 (30.3, 44.8)
Detected with 80% Precision (95% CI)	1.1 (0.0, 2.9)	1.7 (0.0, 3.6)	2.7 (0.6, 5.2)	4.3 (1.6, 7.5)	9.7 (5.6, 14.0)	23.3 (16.8, 29.8)

Abbreviations: UD, unplanned dialysis; CI, confidence interval.

Supplemental Table B-15. Cumulative % of unplanned dialysis patients for which an alert was triggered by time period (12-month model; external evaluation).

	Time to Dialysis Event (months)					
	>=15	***>=12***	***>=9***	***>=6***	***>=3***	***>=0***
Total UD Patients (95% CI)	50.2 (42.8, 57.7)	57.7 (50.6, 64.6)	65.9 (58.9, 72.6)	75.6 (69.2, 81.4)	88.0 (83.2, 92.5)	100.0 (100.0, 100.0)
Detected with 60% Precision (95% CI)	18.6 (13.1, 24.9)	26.2 (19.5, 32.6)	33.8 (26.9, 40.6)	44.7 (37.1, 52.2)	56.0 (48.8, 63.6)	75.1 (68.9, 81.0)
Detected with 70% Precision (95% CI)	12.1 (7.8, 17.1)	16.4 (11.3, 22.1)	21.9 (15.9, 27.9)	30.7 (23.6, 37.8)	44.7 (37.2, 52.6)	61.1 (53.8, 68.5)
Detected with 80% Precision (95% CI)	6.0 (3.1, 9.9)	7.6 (4.1, 11.7)	13.7 (8.7, 18.9)	19.1 (13.3, 24.9)	29.4 (23.0, 36.0)	47.5 (40.0, 54.5)

Abbreviations: UD, unplanned dialysis; CI, confidence interval.

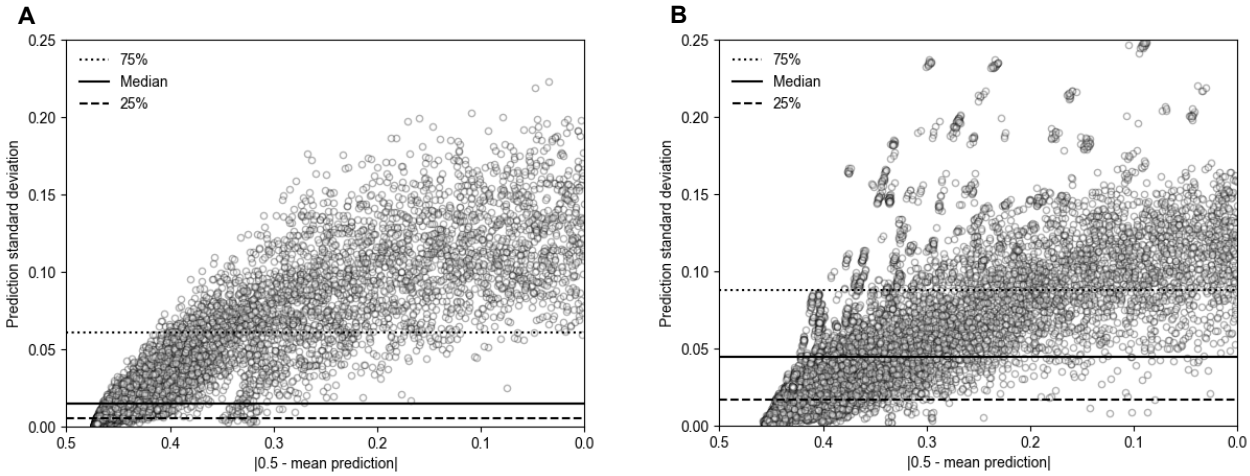
Supplemental Table B-16. Confusion tables obtained from external validation (6-month and 12-month model).

6-Month (70% Precision [a])			
		Predicted	
		False	**True**
Actual	**False**	8597 TN	405 FP
	True	931 FN	657 TP

12-Month (70% Precision [a])			
		Predicted	
		False	**True**
Actual	**False**	6026 TN	866 FP
	True	937 FN	2091 TP

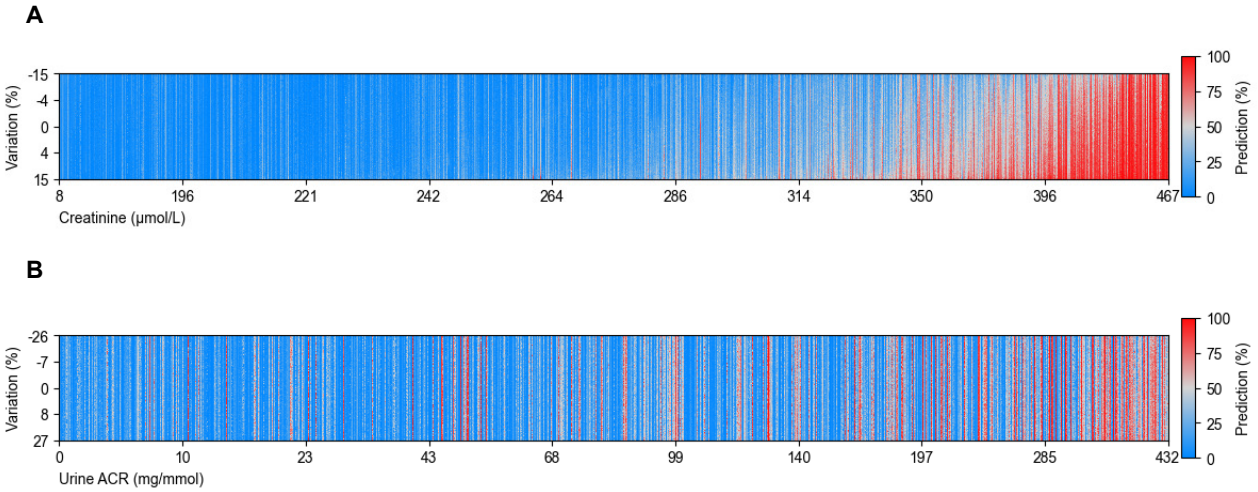
a: Threshold is determined from internal validation results.

Supplemental Figure B-1. Visualization of prediction variability as a function of model confidence.



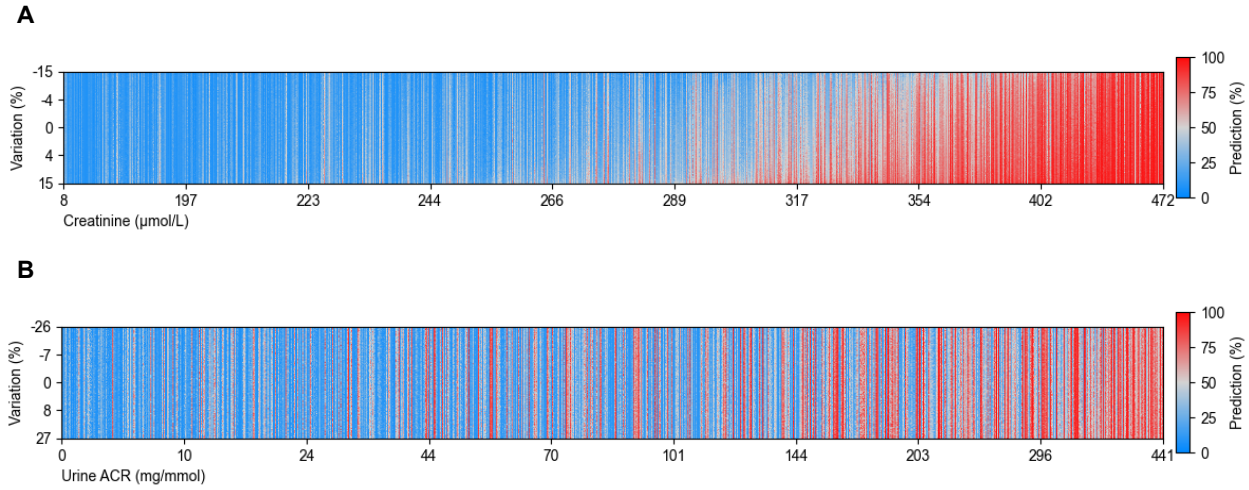
(**A**) 6-month model, and (**B**) 12-month model prediction variability as a function of model confidence. The mean and standard deviation of predictions obtained over 1,000 Monte Carlo samples for each of the examples in the hold-out test set is plotted. The standard deviations are plotted against the model confidence, defined as the absolute of the difference between 0.5 (50%) and the model prediction. Annotated are the 75, 50 (median), and 25 percentiles of the distribution.

Supplemental Figure B-2. Visualization of prediction variability for selected laboratory features (6-month model).



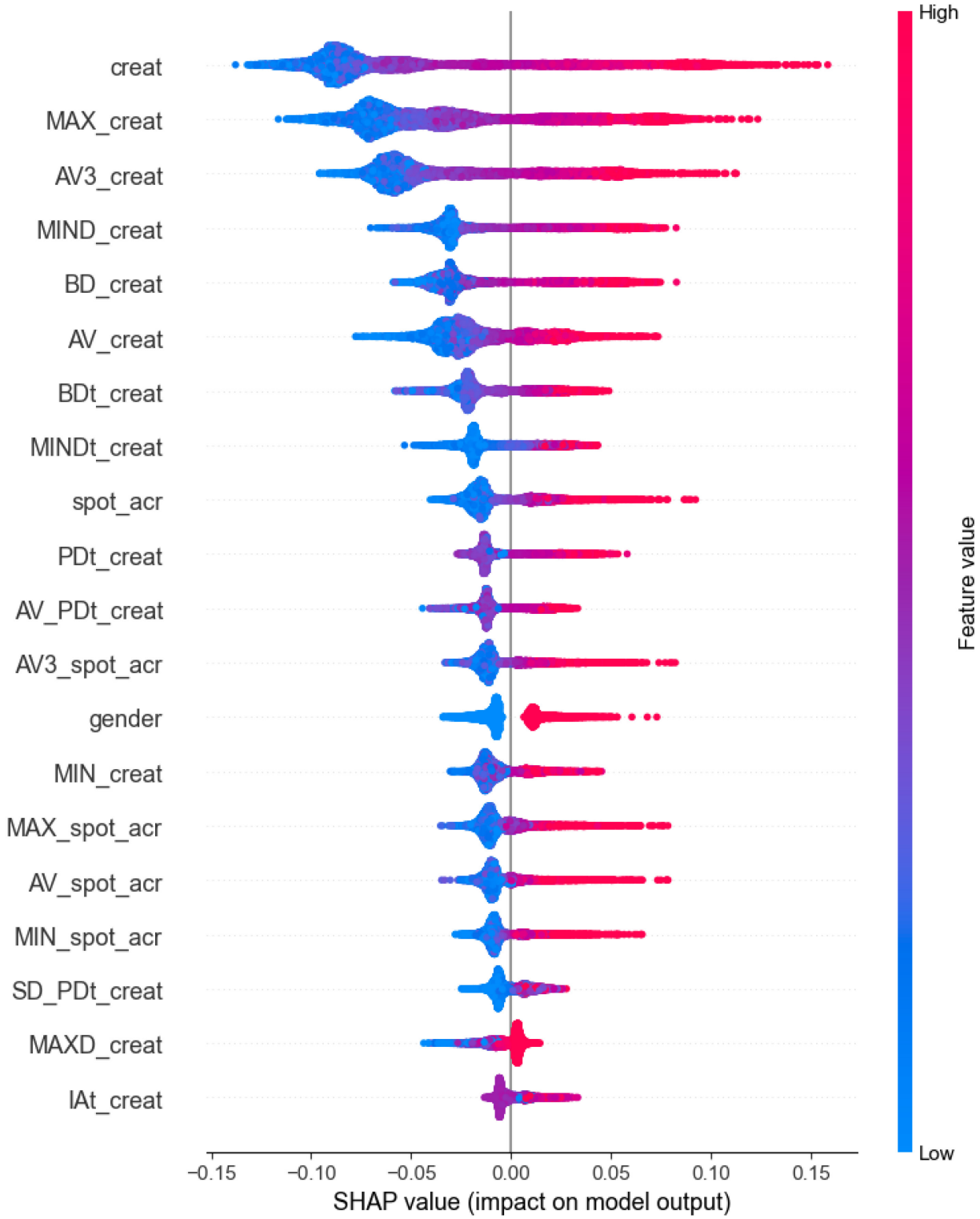
For each example (follow-up visit) in the hold-out test set, laboratory features were randomly perturbed with noise using Monte Carlo sampling. The procedure was repeated for both of the selected laboratory features, (**A**) creatinine, and (**B**) urine albumin-to-creatinine ratio (ACR). Each column in either of the figures represents 1,000 perturbations of the same test example. Within each column, data are sorted according to the applied noise, and in certain cases show how the applied variation can shift a prediction from positive (red), to negative (blue), and vice-versa. Each example (column) is sorted in the x-axis according to the original laboratory value associated with that example, showing how different ranges of laboratory values correlate with model output.

Supplemental Figure B-3. Visualization of prediction variability for selected laboratory features (12-month model).

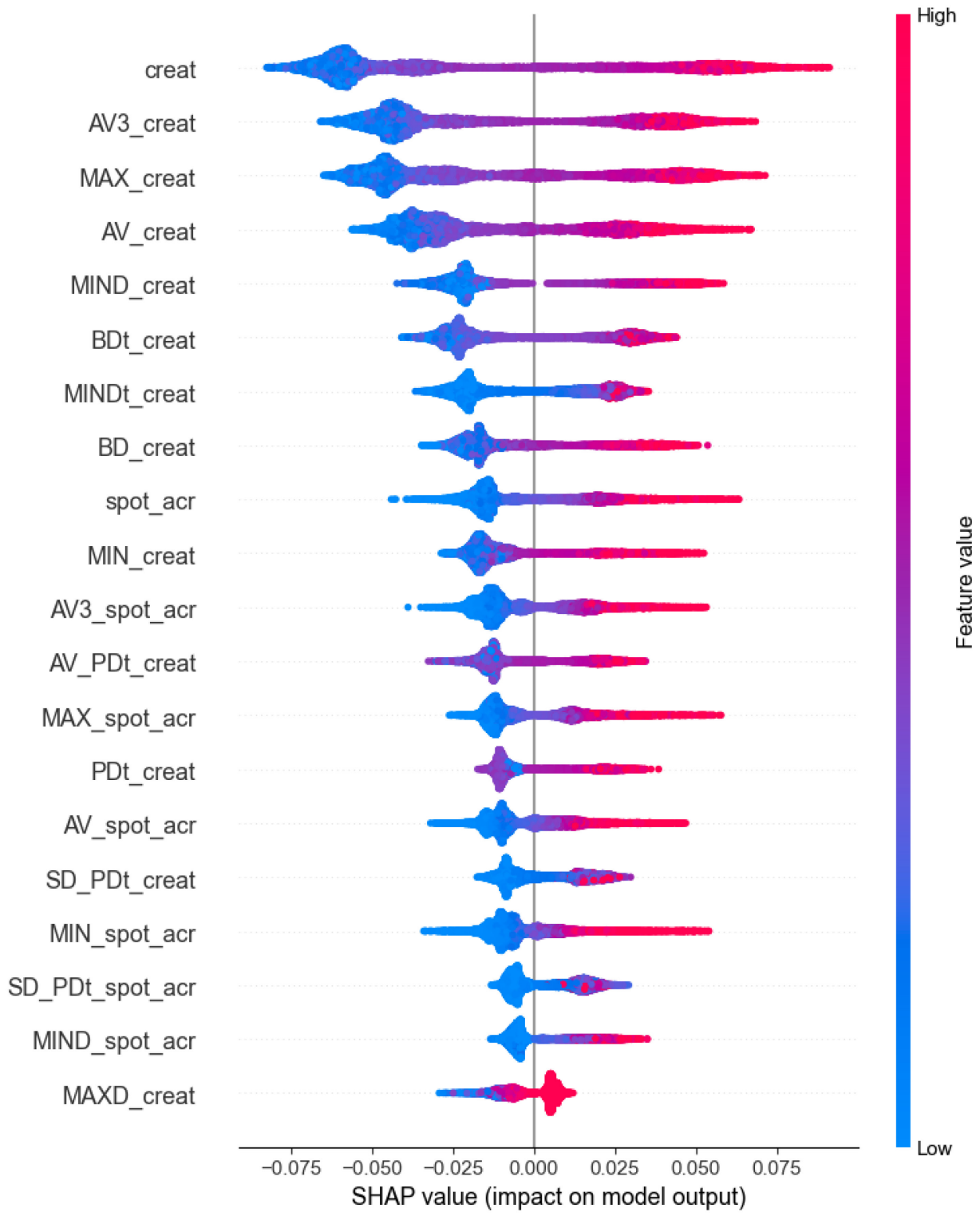


For each example (follow-up visit) in the hold-out test set, laboratory features were randomly perturbed with noise using Monte Carlo sampling. The procedure was repeated for both of the selected laboratory features, (**A**) creatinine, and (**B**) urine albumin-to-creatinine ratio (ACR). Each column in either of the figures represents 1,000 perturbations of the same test example. Within each column, data are sorted according to the applied noise, and in certain cases show how the applied variation can shift a prediction from positive (red), to negative (blue), and vice-versa. Each example (column) is sorted in the x-axis according to the original laboratory value associated with that example, showing how different ranges of laboratory values correlate with model output.

Supplemental Figure B-4. SHAP summary plot (6-month model).



Supplemental Figure B-5. SHAP summary plot (12-month model).



References

1. Collaborators, G.B.D.R.F., *Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1223-1249.
2. McCullough, K.P., et al., *Projecting ESRD Incidence and Prevalence in the United States through 2030.* J Am Soc Nephrol, 2019. **30**(1): p. 127-135.
3. Lorenzo, V., et al., *Predialysis nephrologic care and a functioning arteriovenous fistula at entry are associated with better survival in incident hemodialysis patients: an observational cohort study.* Am J Kidney Dis, 2004. **43**(6): p. 999-1007.
4. Arulkumaran, N., et al., *Causes and risk factors for acute dialysis initiation among patients with end-stage kidney disease-a large retrospective observational cohort study.* Clin Kidney J, 2019. **12**(4): p. 550-558.
5. Brown, P.A., et al., *Factors Associated with Unplanned Dialysis Starts in Patients followed by Nephrologists: A Retropective Cohort Study.* PLoS One, 2015. **10**(6): p. e0130080.
6. Hassan, R., et al., *Risk Factors for Unplanned Dialysis Initiation: A Systematic Review of the Literature.* Canadian Journal of Kidney Health and Disease, 2019. **6**: p. 205435811983168.
7. Machowska, A., et al., *Factors influencing access to education, decision making, and receipt of preferred dialysis modality in unplanned dialysis start patients.* Patient Preference and Adherence, 2016. **Volume 10**: p. 2229-2237.
8. Mendelssohn, D.C., C. Malmberg, and B. Hamandi, *An integrated review of "unplanned" dialysis initiation: reframing the terminology to "suboptimal" initiation.* BMC Nephrol, 2009. **10**: p. 22.
9. Ramspek, C.L., et al., *Kidney Failure Prediction Models: A Comprehensive External Validation Study in Patients with Advanced CKD.* J Am Soc Nephrol, 2021. **32**(5): p. 1174-1186.
10. Tangri, N., et al., *A predictive model for progression of chronic kidney disease to kidney failure.* JAMA, 2011. **305**(15): p. 1553-9.
11. Grams, M.E., et al., *Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate.* Kidney Int, 2018. **93**(6): p. 1442-1451.
12. Drawz, P.E., et al., *A Simple Tool to Predict End-Stage Renal Disease within 1 Year in Elderly Adults with Advanced Chronic Kidney Disease.* Journal of the American Geriatrics Society, 2013. **61**(5): p. 762-768.
13. Hod, T., et al., *Arteriovenous Fistula Placement in the Elderly: When Is the Optimal Time?* Journal of the American Society of Nephrology, 2015. **26**(2): p. 448-456.
14. Kimmel, P.L. and M.E. Rosenberg, *Chronic renal disease*. Second edition. ed. 2020, London: Academic Press/Elsevier. xxi, 1367 pages.
15. *KDIGO Guidelines*. 23/05/2023]; Available from: https://kdigo.org/guidelines/.

16. DALYs, G.B.D. and H. Collaborators, *Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015.* Lancet, 2016. **388**(10053): p. 1603-1658.
17. Manns, B., et al., *The Financial Impact of Advanced Kidney Disease on Canada Pension Plan and Private Disability Insurance Costs.* Canadian Journal of Kidney Health and Disease, 2017. **4**: p. 205435811770398.
18. Arora, P., et al., *Prevalence estimates of chronic kidney disease in Canada: results of a nationally representative survey.* Canadian Medical Association Journal, 2013. **185**(9): p. E417-E423.
19. *Annual statistics on organ replacement in Canada, 2012 to 2021*. Canadian Institute for Health Information.
20. Wuttke, M., et al., *A catalog of genetic loci associated with kidney function from analyses of a million individuals.* Nature Genetics, 2019. **51**(6): p. 957-972.
21. Furman, D., et al., *Chronic inflammation in the etiology of disease across the life span.* Nature Medicine, 2019. **25**(12): p. 1822-1832.
22. Levin, A., et al., *Biomarkers of inflammation, fibrosis, cardiac stretch and injury predict death but not renal replacement therapy at 1 year in a Canadian chronic kidney disease cohort.* Nephrology Dialysis Transplantation, 2014. **29**(5): p. 1037-1047.
23. Rayego-Mateos, S., et al., *Targeting inflammation to treat diabetic kidney disease: the road to 2030.* Kidney Int, 2023. **103**(2): p. 282-296.
24. Ebert, T., et al., *Inflammation and Oxidative Stress in Chronic Kidney Disease and Dialysis Patients.* Antioxidants & Redox Signaling, 2021. **35**(17): p. 1426-1448.
25. Yun, H.R., et al., *Obesity, Metabolic Abnormality, and Progression of CKD.* Am J Kidney Dis, 2018. **72**(3): p. 400-410.
26. Romagnani, P., et al., *Chronic kidney disease.* Nature Reviews Disease Primers, 2017. **3**(1): p. 17088.
27. Inker, L.A., et al., *New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race.* New England Journal of Medicine, 2021. **385**(19): p. 1737-1749.
28. Inker, L.A. and A. Okparavero, *Cystatin C as a marker of glomerular filtration rate: prospects and limitations.* Curr Opin Nephrol Hypertens, 2011. **20**(6): p. 631-9.
29. Inker, L.A., et al., *Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C.* New England Journal of Medicine, 2012. **367**(1): p. 20-29.
30. Robin Heckenauer, J.W., Cédric Wemmert, Friedrich Feuerhake, Michel Hassenforder, et al., *Détection en temps réel des glomérules en pathologie rénale*, in *ORASIS 2021, Centre National de la Recherche Scientifique [CNRS]*. 2021: Saint Ferréol, France.
31. Leung, G., et al., *Could MRI Be Used To Image Kidney Fibrosis? A Review of Recent Advances and Remaining Barriers.* Clin J Am Soc Nephrol, 2017. **12**(6): p. 1019-1028.
32. Field, M.J., C. Pollock, and D. Harris, *The Renal System : Systems of the Body Series*. 2nd ed. Systems of the Body. 2011, London: Elsevier Health Sciences UK.

33. Kidney Disease: Improving Global Outcomes Diabetes Work, G., *KDIGO 2022 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease.* Kidney Int, 2022. **102**(5S): p. S1-S127.
34. Delanaye, P., et al., *Chapter 4 - Assessing Kidney Function*, in *Chronic Renal Disease (Second Edition)*, P.L. Kimmel and M.E. Rosenberg, Editors. 2020, Academic Press. p. 37-54.
35. Hojs, R., et al., *Kidney function estimating equations in patients with chronic kidney disease.* International Journal of Clinical Practice, 2011. **65**(4): p. 458-464.
36. Levey, A.S., et al., *A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group.* Ann Intern Med, 1999. **130**(6): p. 461-70.
37. Levey, A.S., et al., *A new equation to estimate glomerular filtration rate.* Ann Intern Med, 2009. **150**(9): p. 604-12.
38. Marzinke, M.A., et al., *Limited Evidence for Use of a Black Race Modifier in eGFR Calculations: A Systematic Review.* Clin Chem, 2022. **68**(4): p. 521-533.
39. Zacharias, H.U., et al., *A Predictive Model for Progression of CKD to Kidney Failure Based on Routine Laboratory Tests.* Am J Kidney Dis, 2022. **79**(2): p. 217-230 e1.
40. Ramspek, C.L., et al., *Towards the best kidney failure prediction tool: a systematic review and selection aid.* Nephrology Dialysis Transplantation, 2020. **35**(9): p. 1527-1538.
41. Mazzaferro, S., et al., *Focus on the Possible Role of Dietary Sodium, Potassium, Phosphate, Magnesium, and Calcium on CKD Progression.* Journal of Clinical Medicine, 2021. **10**(5): p. 958.
42. Gluba-Brzózka, A., B. Franczyk, and J. Rysz, *Vegetarian Diet in Chronic Kidney Disease—A Friend or Foe.* Nutrients, 2017. **9**(4): p. 374.
43. Baker, M. and M.A. Perazella, *NSAIDs in CKD: Are They Safe?* American Journal of Kidney Diseases, 2020. **76**(4): p. 546-557.
44. Narva, A.S., J.M. Norton, and L.E. Boulware, *Educating Patients about CKD: The Path to Self-Management and Patient-Centered Care.* Clin J Am Soc Nephrol, 2016. **11**(4): p. 694-703.
45. Taal, M.W. and B.M. Brenner, *Predicting initiation and progression of chronic kidney disease: Developing renal risk scores.* Kidney International, 2006. **70**(10): p. 1694-1705.
46. Landray, M.J., et al., *Prediction of ESRD and death among people with CKD: the Chronic Renal Impairment in Birmingham (CRIB) prospective cohort study.* Am J Kidney Dis, 2010. **56**(6): p. 1082-94.
47. Al-Wahsh, H., et al., *Accounting for the Competing Risk of Death to Predict Kidney Failure in Adults With Stage 4 Chronic Kidney Disease.* JAMA Netw Open, 2021. **4**(5): p. e219225.
48. Ferguson, T., et al., *Development and External Validation of a Machine Learning Model for Progression of CKD.* Kidney International Reports, 2022.
49. Tangri, N., et al., *Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis.* JAMA, 2016. **315**(2): p. 164-74.

50. Harasemiw, O., et al., *Integrating Risk-Based Care for Patients With Chronic Kidney Disease in the Community: Study Protocol for a Cluster Randomized Trial.* Canadian Journal of Kidney Health and Disease, 2019. **6**: p. 205435811984161.
51. Collett, D., *Modelling survival data in medical research*. Third edition. ed. Texts in statistical science. 2015: CRC Press. xvi, 532 pages.
52. Singer, J.D. and J.B. Willett, *Applied longitudinal data analysis : modeling change and event occurrence*. 2003, New York ; Oxford: Oxford University Press.
53. John, D.K. and L.P. Ross, *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Vol. 2nd ed. 2002, Hoboken, N.J.: Wiley-Interscience.
54. Walker, S.R., et al., *Association of frailty and physical function in patients with non-dialysis CKD: a systematic review.* BMC Nephrology, 2013. **14**(1): p. 228.
55. Hundemer, G.L., et al., *Social determinants of health and the transition from advanced chronic kidney disease to kidney failure.* Nephrol Dial Transplant, 2023. **38**(7): p. 1682-1690.
56. Barraclough, H., L. Simms, and R. Govindan, *Biostatistics Primer: What a Clinician Ought to Know: Hazard Ratios.* Journal of Thoracic Oncology, 2011. **6**(6): p. 978-982.
57. Altman, D.G. and B.L. De Stavola, *Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates.* Statistics in Medicine, 1994. **13**(4): p. 301-341.
58. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. Adaptive computation and machine learning. 2016, Cambridge, Massachusetts: The MIT Press.
59. Grinsztajn, L., E. Oyallon, and G. Varoquaux, *Why do tree-based models still outperform deep learning on tabular data?* arXiv pre-print server, 2022.
60. Vickers, A.J. and A.M. Cronin, *Traditional Statistical Methods for Evaluating Prediction Models Are Uninformative as to Clinical Value: Towards a Decision Analytic Framework.* Seminars in Oncology, 2010. **37**(1): p. 31-38.
61. Tin Kam, H. *Random decision forests*. in *Proceedings of 3rd International Conference on Document Analysis and Recognition*. 1995.
62. Breiman, L., *Classification and regression trees*. 1998, Boca Raton: Chapman & Hall/CRC.
63. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825--2830.
64. Hyafil, L. and R.L. Rivest, *Constructing optimal binary decision trees is NP-complete.* Information Processing Letters, 1976. **5**(1): p. 15-17.
65. Ishwaran, H., et al., *Random survival forests.* 2008.
66. Martin, K.J. and E.A. Gonzalez, *Metabolic bone disease in chronic kidney disease.* J Am Soc Nephrol, 2007. **18**(3): p. 875-85.
67. Ramspek, C.L., et al., *Predicting Kidney Failure, Cardiovascular Disease and Death in Advanced CKD Patients.* Kidney International Reports, 2022. **7**(10): p. 2230-2241.

68. Ramspek, C.L., et al., *Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models.* Int J Epidemiol, 2021.
69. van Walraven, C., C. McCudden, and P.C. Austin, *Imputing missing laboratory results may return erroneous values because they are not missing at random.* J Clin Epidemiol, 2023. **154**: p. 65-74.
70. Moritz, S., et al., *Comparison of different Methods for Univariate Time Series Imputation in R.* arXiv pre-print server, 2015.
71. McCudden, C., et al., *Individual patient variability with the application of the kidney failure risk equation in advanced chronic kidney disease.* PLOS ONE, 2018. **13**(6): p. e0198456.
72. Tangri, N., et al., *A Dynamic Predictive Model for Progression of CKD.* Am J Kidney Dis, 2017. **69**(4): p. 514-520.
73. Tomašev, N., et al., *A clinically applicable approach to continuous prediction of future acute kidney injury.* Nature, 2019. **572**(7767): p. 116-119.
74. Tomasev, N., et al., *Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records.* Nat Protoc, 2021. **16**(6): p. 2765-2787.
75. Burkov, A., *Machine learning engineering*. 2020, Place of publication not identified: True Positive, Inc. xxvi, 282 pages : illustrations (some colour).
76. Norgeot, B., et al., *Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist.* Nat Med, 2020. **26**(9): p. 1320-1324.
77. Drawz, P.E., et al., *A simple tool to predict end-stage renal disease within 1 year in elderly adults with advanced chronic kidney disease.* J Am Geriatr Soc, 2013. **61**(5): p. 762-8.
78. Vickers, A.J. and A.M. Cronin, *Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework.* Semin Oncol, 2010. **37**(1): p. 31-8.
79. Shapley, L.S., *A value for n-person games*. 1952, Santa Monica, Calif.,: Rand Corp. 13 l.
80. Štrumbelj, E. and I. Kononenko, *Explaining prediction models and individual predictions with feature contributions.* Knowledge and Information Systems, 2014. **41**(3): p. 647-665.
81. Hod, T., et al., *Arteriovenous fistula placement in the elderly: when is the optimal time?* J Am Soc Nephrol, 2015. **26**(2): p. 448-56.
82. Buck, J., et al., *Why do patients known to renal services still undergo urgent dialysis initiation? A cross-sectional survey.* Nephrol Dial Transplant, 2007. **22**(11): p. 3240-5.
83. Chiu, K., A. Alam, and S. Iqbal, *Predictors of suboptimal and crash initiation of dialysis at two tertiary care centers.* Hemodial Int, 2012. **16 Suppl 1**: p. S39-46.
84. Crews, D.C., et al., *Inpatient hemodialysis initiation: reasons, risk factors and outcomes.* Nephron Clin Pract, 2010. **114**(1): p. c19-28.
85. Holland, D.C. and M. Lam, *Suboptimal dialysis initiation in a retrospective cohort of predialysis patients--predictors of in-hospital dialysis initiation, catheter insertion and one-year mortality.* Scand J Urol Nephrol, 2000. **34**(6): p. 341-7.

86. Hughes, S.A., et al., *Factors associated with suboptimal initiation of dialysis despite early nephrologist referral.* Nephrol Dial Transplant, 2013. **28**(2): p. 392-7.
87. Mendelssohn, D.C., et al., *Suboptimal initiation of dialysis with and without early referral to a nephrologist.* Nephrol Dial Transplant, 2011. **26**(9): p. 2959-65.
88. System, U.S.R.D., *2021 USRDS Annual Data Report: Epidemiology of kidney disease in the United States*. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.
89. Caskey, F.J., et al., *Early referral and planned initiation of dialysis: what impact on quality of life?* Nephrol Dial Transplant, 2003. **18**(7): p. 1330-8.
90. Green, D., et al., *How accurately do nephrologists predict the need for dialysis within one year?* Nephron Clin Pract, 2012. **122**(3-4): p. 102-6.
91. Ferguson, T., et al., *Development and External Validation of a Machine Learning Model for Progression of CKD.* Kidney Int Rep, 2022. **7**(8): p. 1772-1781.
92. Rabbani, N., et al., *Applications of machine learning in routine laboratory medicine: Current state and future directions.* Clin Biochem, 2022. **103**: p. 1-7.
93. Sylvestre, M.-P., *traj: Trajectory Analysis*. 2023.
94. Van Den Brand, J.A.J.G., et al., *Predicting kidney failure from longitudinal kidney function trajectory: A comparison of models.* PLOS ONE, 2019. **14**(5): p. e0216559.
95. Shah, B.V. and A.S. Levey, *Spontaneous changes in the rate of decline in reciprocal serum creatinine: errors in predicting the progression of renal disease from extrapolation of the slope.* Journal of the American Society of Nephrology, 1992. **2**(7): p. 1186-1191.
96. Tin Kam, H. *Random decision forests*. IEEE Comput. Soc. Press.
97. Forman, G. and M. Scholz, *Apples-to-apples in cross-validation studies.* ACM SIGKDD Explorations Newsletter, 2010. **12**(1): p. 49-57.
98. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.
99. Slinin, Y., et al., *Timing of dialysis initiation, duration and frequency of hemodialysis sessions, and membrane flux: a systematic review for a KDOQI clinical practice guideline.* Am J Kidney Dis, 2015. **66**(5): p. 823-36.
100. Cooper, B.A., et al., *A Randomized, Controlled Trial of Early versus Late Initiation of Dialysis.* New England Journal of Medicine, 2010. **363**(7): p. 609-619.
101. Cao, J., et al., *Generalizability of an acute kidney injury prediction model across health systems.* Nature Machine Intelligence, 2022. **4**(12): p. 1121-1129.
102. Bansal, N., et al., *Development and validation of a model to predict 5-year risk of death without ESRD among older adults with CKD.* Clin J Am Soc Nephrol, 2015. **10**(3): p. 363-71.
103. Hundemer, G.L., et al., *The Effect of Age on Performance of the Kidney Failure Risk Equation in Advanced CKD.* Kidney Int Rep, 2021. **6**(12): p. 2993-3001.
104. Rosansky, S.J. and R.J. Glassock, *Is a decline in estimated GFR an appropriate surrogate end point for renoprotection drug trials?* Kidney International, 2014. **85**(4): p. 723-727.

105. Badve, S.V., et al., *Glomerular filtration rate decline as a surrogate end point in kidney disease progression trials.* Nephrology Dialysis Transplantation, 2015. **31**(9): p. 1425-1436.
106. Inker, L.A., et al., *New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race.* N Engl J Med, 2021. **385**(19): p. 1737-1749.
107. Moncada-Torres, A., et al., *Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival.* Scientific Reports, 2021. **11**(1).
108. Géron, A.l., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Second edition. ed. 2019, Beijing China ; Sebastopol, CA: O'Reilly Media, Inc. xxv, 819 pages.
109. Davidson-Pilon, C., *lifelines: survival analysis in Python.* Journal of Open Source Software, 2019. **4**(40): p. 1317.
110. Pölsterl, S., *scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn.* Journal of Machine Learning Research, 2020. **21**(212): p. 1-6.
111. Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.
112. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.