KU LEUVEN

DOCTORAL SCHOOL OF BIOMEDICAL SCIENCES
Faculty of Medicine

Development of predictive models for critically ill patients with acute kidney injury

Fateme Nateghi Haredasht

Supervisors:
Prof. dr. Celine Vens
Prof. dr. Hans Pottel
dr. Wouter De Corte
dr. Liesbeth Viaene

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Biomedical Sciences

January 2023

KU Leuven

Biomedical Sciences Group

Faculty of Medicine

Department of Public Health and Primary Care

KU LEUVEN

DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

Development of predictive models for critically ill patients with acute kidney injury

Fateme NATEGHI HAREDASHT

Examination committee:
Prof. dr. ir. Simon De Meyer, chair
Prof. dr. Celine Vens, supervisor
Prof. dr. Hans Pottel, supervisor
dr. Wouter De Corte, supervisor
dr. Liesbeth Viaene, supervisor
Prof. dr. Geert Molenberghs
Prof. dr. Geert Meyfroidt
Prof. dr. Pierre Delanaye
(University of Liège)
Prof. dr. Isaac Triguero Velázquez
(University of Nottingham)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Biomedical Sciences

January 2023

Uitgegeven in eigen beheer, Fateme Nateghi Haredasht, Oude Markt 13, 3000 Leuven (Belgium)

Acknowledgements

The completion of this thesis would not have been possible without the help and guidance of many people. It is therefore my pleasure to express my gratitude to everyone who has contributed to this work. In the first place, I would like to express my gratitude and thanks to my thesis supervisors. Celine, having you as a promoter was an honor and a privilege. You have been an excellent guide to me throughout my PhD journey, and I am extremely grateful for this. I can create a long list of things that I learned from you. Your insights helped me see the bigger picture and take a fresh approach to our work. Due to the amazingly friendly and international lab that you created, I never felt far from home. Thank you for everything!

Hans, it has been a true honor to conduct my PhD under your supervision and to have such a talented and visionary statistician as a co-promoter. You enthusiastically shared your statistical knowledge and expertise, which helped me feel at ease with our projects.

Wouter and Liesbeth, I could not have wished to have better clinicians as co-promoters than you. You both kindly shared your research and clinical expertise and made me understand the clinical impact of our work. I very much enjoyed coming to the AZ Groeninge hospital during the first year of my PhD which I gained the ability to interact with a diverse team, including clinicians, IT staff, and scientists across domains. Also, I want to thank Christophe for his patience and helpful assistance during my hospital visit.

I am also grateful to my Jury members. My Thesis Advisory Committee members, Professor Geert Molenberghs and Professor Geert Meyfroidt thank you so much for your enthusiasm and for following my work over the past four years. Professor Pierre Delanaye, I had the pleasure to collaborate with you on a project. Thank you for agreeing to be part of my jury committee and for reviewing my thesis. Professor Isaac Triguero Velázquez, thank you for your interest in our research and your valuable feedback. Finally, I would like to thank Professor Simon De Meyer, for chairing the public defense.

I am also very grateful to all my wonderfully kind and amazing colleagues in our research lab. Konstantinos, Alireza, Jasper, Felipe, Klest, Michela, and Robbe thank you for all the great memories that I have from our social activities and research collaboration. Jasper and Felipe, although most of our time together was spent at Covid, I enjoyed being your officemate. Robbe, it was always fun talking with you about our favorite tea types! Thank you, Klest, for organizing the activities for internationals in Kulak. Thank you both Konstantinos and Alireza for answering my questions and providing feedback for my PhD milestones. A particular thanks to Michela for being an amazing and supportive friend and a talented collaborator on additional research topics together. Thanks a million to all of you!

I would also like to state my gratitude towards ITEC-imec group, Piet, Ine, Frederik, and other members. Being part of a research group that is as dynamic and versatile as yours was a pleasure for me. Furthermore, I am thankful to the Doctoral Working Group and the international group of KULAK for the brilliant events and the great moments we had together. I am also thankful to everyone from the faculty of medicine who were so kind to me, Melanie, Petra, Elien, Timo, and others.

My deepest acknowledgment goes to my parents and brothers, as I would not be standing here today without their support over the years. I am grateful for your trust in me and the freedom you provided me to pursue my interests. I would also like to thank my parents-in-law for their kindness and support. Also a huge thanks to my Iranian friends in Gent for their presence and kindness during these years.

Last but not least, I would like to thank my dearest Pooya. I find it difficult to express my appreciation because it is so boundless. You are my best friend and my most enthusiastic cheerleader. As the wife of a science enthusiast, I am extremely grateful to you. Whether it is a scientific problem or a personal issue, you are always there for me. There are no words to describe how grateful I am to be with someone I have known for more than 15 years. I love you!

Abstract

During the past few years, artificial intelligence (AI) and machine learning has become a huge trend in different domains. AI is the simulation of human intelligence processes by machines, especially computers, and machine learning is defined as a discipline of AI that provides machines the ability to automatically learn. In various applications such as finance and healthcare, machine learning is omnipresent and widely used. The recent digitization of health records has provided an excellent environment for assessing the usability of such techniques in healthcare. In consequence, the field is now seeing an increase in research papers involving machine learning applied to electronic health records (EHR) for the purpose of accurate understanding from historical data (descriptive analysis) as well as predicting health risks and health trajectories (predictive analysis). In many clinical studies, the outcome of interest is the time until some event occurs in which such time-to-event data is also called survival data. It is surprising that survival data has not received much attention from the research community of data mining and machine learning. The lack of attention may be due to the fact that standard machine learning techniques cannot be applied directly to survival data, primarily due to censoring. The fundamental contributions of this PhD project include the adaptation of existing and the development of new machine learning techniques. The project objective can be separated into methodological objectives, which make a substantial contribution to the field of machine learning, and medical application objectives, which will lead to an improved post-ICU policy for acute kidney injury (AKI) patients. More specifically, we focus on developing new machine learning-based methods to predict time-to-event and also we address the societal and economic challenge posed by ICU-related acute kidney injury (AKI) by employing machine learning models to predict the risk of chronic kidney disease (CKD) for AKI survivors.

As there are multiple definitions of AKI, analyzing its incidence and associated outcomes is challenging. We first started our work by investigating the effect of different existing definitions of AKI on predicting adverse outcomes (in-hospital mortality). To identify AKI patients at risk of in-hospital mortality, we

employed a machine-learning model. Then we continued our work by conducting a systematic review to find existing validated risk prediction models (statistical or machine learning) for outcomes of AKI. We realized that state-of-the-art machine learning models using data information are required to increase the predictive performance for developing renal insufficiencies after AKI.

Many clinical time-to-event studies that require follow-up of the patients after a hospital stay form a logistic challenge because once the patients take up their normal activities, it is often difficult to reach or motivate them to continue their participation in the study. This results in a high rate of drop-out, and for many patients, no follow-up data is available beyond that of their hospital stay. Nevertheless, the training set can be easily augmented with the retrospective hospital data from patients who are not participating in the study in many of these prospective studies. In either case, if the study outcome is determined during follow-up, we have a substantial part of the training set that is unlabeled (equivalently, the censoring time for these patients is zero). AI and machine learning models are only as good as the data used to train them. Over the past few years, machine learning-based techniques have become increasingly popular in survival analysis. However, applying machine learning methods directly to censored data is challenging since the value of a measurement or observation is only partially known. We modified a semi-supervised learning algorithm to make use of censored information in survival analysis. We first proposed a novel idea augmenting the training set with the unlabeled data using the self-training algorithm. We developed three approaches of which two were based on a semi-supervised algorithm (self-training) for augmenting the labeled set.

We continued our work by developing a new machine learning-based model for predicting time-to-event. The proposed model transforms the time-to-event prediction problem into a semi-supervised regression problem. In the proposed approach, called STUART, censored observations were introduced as partially labeled observations since their target values should exceed the censoring time.

As our main objective, we have employed machine-learning-based models to predict outcomes following severe AKI events in the ICU. Our work supported the view that machine learning-based models have the potential to help clinical decision-making for identifying those patients that have a higher chance of developing CKD after hospital discharge from critically ill patients who experienced severe AKI. Also, we verified our novel idea that the inclusion of unlabeled data points in the survival analysis task results in achieving a better predictive performance in a survival prediction task.

The impact of this work is considered to be two-fold. First, in the area of healthcare since we believe our work will lead to an improved post-ICU policy

for AKI patients. Second, in the area of machine learning itself since new learning algorithms for time-to-event data are delivered.

Beknopte samenvatting

In de afgelopen jaren is machine learning een enorme trend in de industrie geworden. In verschillende toepassingen, zoals financiën en gezondheidszorg, is machine learning alomtegenwoordig en veel gebruikt. De recente digitalisatie van medische dossiers heeft een uitstekende omgeving gecreëerd om de bruikbaarheid van dergelijke technieken in de gezondheidszorg te beoordelen. Als gevolg hiervan zijn er steeds meer onderzoekspapers waarbij machine learning wordt toegepast op elektronische medische dossiers om gezondheidsrisico's en gezondheidstrajecten te voorspellen. In veel klinische onderzoeken is men geïnteresseerd in de tijd totdat een gebeurtenis plaatsvindt, waarbij dergelijke gegevens rond tijd-tot-gebeurtenis ook wel overlevingsgegevens genoemd worden. Verrassend genoeg hebben overlevingsgegevens nog maar weinig aandacht gekregen van de onderzoeksgemeenschap van data mining en machine learning. Het gebrek aan aandacht kan te wijten zijn aan het feit dat standaard technieken voor machine learning niet direct kunnen worden toegepast op overlevingsgegevens, voornamelijk vanwege gecensureerde observaties. De fundamentele bijdragen van dit doctoraatsproject omvatten de adoptie van bestaande en de ontwikkeling van nieuwe technieken voor machine learning. De projectdoelstelling kan worden opgesplitst in methodologische doelstellingen, die een substantiële bijdrage leveren aan het veld van machine learning, en medische toepassingsdoelstellingen, die zullen leiden tot een verbeterd post-intensieve zorg (IZ) beleid voor AKI-patiënten. Meer specifiek richten we ons op het ontwikkelen van nieuwe machine learning methoden om de tijd tot een gebeurtenis te voorspellen en we pakken ook de maatschappelijke en economische uitdaging aan van IZ-gerelateerd acute nierschade (AKI) door machine learning-modellen te gebruiken om het risico op chronische nierziekte (CKD) voor AKI-overlevenden te voorspellen. Aangezien er meerdere definities van AKI zijn, is het analyseren van de incidentie en de bijhorende gevolgen een uitdaging. We begonnen ons werk met het onderzoeken van het effect van verschillende bestaande definities van AKI op het voorspellen van ongunstige gevolgen (mortaliteit in het ziekenhuis). Om AKI-patienten met een risico

op ziekenhuissterfte te identificeren, hebben we een machine learning-model gebruikt. Vervolgens hebben we ons werk voortgezet met een systematische review van de bestaande gevalideerde risicovoorspellingsmodellen (statistisch of machine learning) voor gevolgen van AKI. We realiseerden ons dat geavanceerde machine learning-modellen, die gebruikmaken van big data, nodig zijn om de voorspellingsprestaties voor het ontwikkelen van nierinsufficiënties na AKI te verbeteren.

Veel klinische tijd-tot-gebeurtenis studies die een follow-up van de patiënten na een ziekenhuisopname vereisen, vormen een logistieke uitdaging omdat, wanneer de patiënten eenmaal hun normale activiteiten hebben hervat, het vaak moeilijk is om hen te bereiken of te motiveren om door te gaan met hun deelname aan het onderzoek . Dit resulteert in een hoog uitvalpercentage en voor veel patiënten zijn er zelfs geen follow-upgegevens beschikbaar buiten die van hun ziekenhuisverblijf. Bovendien kunnen retrospectieve ziekenhuisgegevens van patiënten die niet deelnemen aan het onderzoek eenvoudig worden aangevuld met de trainingsset in veel van deze prospectieve studies. In beide gevallen, als het resultaat van de studie wordt bepaald tijdens de follow-up, hebben we een aanzienlijk deel van de trainingsset dat niet-gelabeld is (met andere woorden, de censureringstijd voor deze patiënten is nul). AI- en machine learning-modellen zijn slechts zo goed als de data die worden gebruikt om ze te trainen. We hebben eerst een nieuw idee voorgesteld om deze niet-gelabelde gegevens te verrijken met de trainingsset met behulp van een semi-gesuperviseerd leeralgoritme. We hebben drie methoden ontwikkeld, waarvan er twee gebaseerd waren op een semi-gesuperviseerd algoritme (zelftraining) voor het verrijken van de ongelabelde set. We hebben ons werk voortgezet met de ontwikkeling van een nieuw, op machine learning gebaseerd model voor het voorspellen van de tijd-tot-gebeurtenis. Het voorgestelde model transformeert het probleem van het voorspellen van de tijd-tot-gebeurtenis in een semi-gesuperviseerd regressieprobleem. In de voorgestelde aanpak, STUART genaamd, werden gecensureerde waarnemingen ingevoerd als gedeeltelijk gelabelde waarnemingen, aangezien hun tijd-tot-gebeurtenis de censureringstijd zouden moeten overschrijden. Als belangrijkste doelstelling hebben we op machine learning gebaseerde modellen gebruikt om de gevolgen te voorspellen van ernstige AKI op de ICU. Ons werk ondersteunde de opvatting dat op machine learning gebaseerde modellen het potentieel hebben om klinische besluitvorming te helpen bij het identificeren van die patiënten die een grotere kans hebben om CKD te ontwikkelen na ontslag uit het ziekenhuis, onder de kritisch zieke patiënten die ernstige AKI hebben doorgemaakt. Ook hebben wij ons nieuwe idee geverifieerd dat het opnemen van ongelabelde datapunten in de overlevingsanalysetaak leidt tot betere voorspellende prestaties in een overlevingsvoorspellingstaak. De impact van dit werk wordt als tweeledig beschouwd. Ten eerste op het gebied van gezondheidszorg, omdat we denken dat ons werk zal leiden tot een verbeterd post-IZ-beleid voor AKI-patiënten. Ten

tweede op het gebied van machine learning zelf, omdat nieuwe leeralgoritmen voor tijd-tot-gebeurtenis gegevens worden geleverd.

List of abbreviations

AFT Accelerated failure time

AI Artificial intelligence

AKD Acute kidney disease

AKI Acute kidney injury

AKIN Acute Kidney Injury Network

ATN Acute renal failure Trial Network

AUROC Area under the receiver operating characteristic curve

ANN Artificial neural networks

CSA-AKI Cardiac surgery-associated acute kidney injury

CVD Cardiovascular disease

CKD Chronic kidney disease

CKD-EPI Chronic Kidney Disease Epidemiology Collaboration

CART Classification and regression tree

CI Confidence intervals

CPH Cox Proportional Hazards model

ANN Artificial neural networks

Lasso-Cox Cox regression with LASSO regularization

CHF Cumulative Hazard Function

CysC Cystatin C

DTs Decision trees

AKI-D Dialysis-requiring adult acute kidney injury

EN Elastic-Net

EHR Electronic health record

ESRD End-stage renal disease

EKFC European Kidney Function Consortium

FN False negative

FP False positive

FPR False positive rate

GFR Glomerular filtration rate

C-index Harrell's concordance index

ICUs Intensive care units

IQR Interquartile ranges

IDMS Isotope dilution mass spectrometric method

KDIGO Kidney Disease: Improving Global Outcomes

Lasso Least absolute shrinkage and selection operator

LoS Length of stay

MeSH Medical Subject Heading

ML Machine learning

NSAIDS Non-steroidal anti-inflammatory drugs

PDMS Patient Data Management System

PRISMA Preferred Reporting Items for Systematic Review and Meta-Analysis

RFs Random forests

RSFs Random survival forests

RNN-LSTM Recurrent neural network-long short-term memory

RRT Renal replacement therapy

RIFLE Risk, Injury, Failure, Loss-of-kidney-function, and End-stage kidney disease

ST-RSF Self-trained random survival forest

ST-RSF+CCT Self-trained random survival forest corrected with censored times

STUART Self-Trained sUrvivAl foResT

SSL Semi-supervised learning

SCr Serum creatinine

SVMs Support vector machines

TN True negative

TP True positive

TPR True positive rate

UO Urine output

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This Ph.D. thesis is focused on developing predictive modeling techniques to tackle a number of machine learning challenges related to data analysis within an intensive care unit (ICU) context. An introduction to the subject matter is provided at the beginning of this thesis. Next, we discuss developing clinical prediction models, explaining the structure of clinical data, analyzing such data, and building predictive models using such data. Finally, we provide an overview of the thesis.

1.1 Acute kidney injury

A large number of healthcare resources are consumed every year by millions of patients suffering from critical illnesses [1]. These patients exhibit a diagnostic-therapeutic cycle that is characterized by rapid changes in clinical conditions. In order to prevent chronic phases of critical illness, dedicated therapies should be administered early to patients most susceptible to specific organ deterioration. As a result, intensive care medicine relies heavily on prediction.

Among critically ill patients in the ICU, acute kidney injury (AKI) is one of the most prevalent conditions. Approximately 40% of critically ill patients in the ICUs are affected by AKI [140]. The overall incidence of AKI in hospital patients ranges between 7% to 22%, and it ranges from 20% to 50% in ICU patients [201, 212, 116, 20]. It has been shown that when sepsis is present at ICU admission, the prevalence of AKI is greater than 40% [4]. Moreover, a cohort analysis reported AKI in 36% of the patients on the day after admission to ICU, and a prevalence of more than 60% during ICU stay [74].

Table 1.1: AKI staging according to KDIGO criteria.

AKI Stages	SCr criteria	Urine volume criteria
Stage 1	1.5–1.9 times baseline OR >= 0.3 mg/dL (>= 26.5 μmol/L) absolute increase	Urine volume <0.5 mL/kg/h for 6–12 hours
Stage 2	>= 2.0˜2.9 times baseline	Urine volume <0.5 mL/kg/h for >= 12 hours
Stage 3	>= 3.0 times baseline OR >= 4.0 mg/dL(>= 353.6 μ mol/L) absolute increase OR Initiation of renal replacement therapy OR, In patients <18 years, decrease in eGFR to <35 mL/min per 1.73 m2	Urine volume <0.3 mL/kg/h for >= 24 hours OR Anuria for >= 12 hours

Multiple etiologies like acute tubular necrosis (ATN), rapidly progressive glomerulonephritis, and interstitial nephritis and risk factors contribute to the pathogenesis of AKI. Risk factors include increasing age, presence of heart failure, liver failure, CKD, anemia, and exposures to nephrotoxic agents including antibiotics, non-steroidal anti-inflammatory drugs (NSAIDS), and radiocontrast dyes [129]. Infections, sepsis, shock, need for mechanical ventilation, and surgery is well recognized as high-risk settings for the development of AKI.

Over time, the definition of AKI has changed. In 2012, the Kidney Disease: Improving Global Outcomes (KDIGO) unified the previous definitions [119]. By KDIGO definition, AKI is diagnosed by an absolute increase in serum creatinine (SCr), at least 0.3 mg/dL 26.5 ($\mu mol/L$) within 48 hours or by a 50% increase in SCr from baseline within 7 days, or a urine volume of less than 0.5 $mL/kg/h$ for at least 6 hours [92]. Table 1.1 shows AKI staging according to KDIGO criteria.

Measurement of glomerular filtration rate (GFR) is widely accepted as the most accurate indicator of renal function in both healthy and diseased individuals. GFR can be measured directly by clearance studies of ideal exogenous markers, such as inulin. However, none of these procedures are practical or economical for routine use and serum levels of endogenous filtration markers have traditionally been used to estimate renal function [151].

Serum creatinine (SCr) is the most widely used endogenous filtration marker for assessing renal function in clinical practice. However, SCr is insensitive to detect early renal disease, and levels could remain within the normal range even when renal function is significantly impaired [192]. For example, changes in muscle mass and protein metabolism significantly affect SCr levels. Guidelines, therefore, recommend the use of prediction equations, such as the Modification of Diet in Renal Disease (MDRD) (=not recommended anymore) [109] and CKD-EPI equation, to estimate GFR (eGFR) whenever SCr is measured ([110]). Despite being used in literature, CKD-EPI equations lack continuity with aging. Pottel et al. [150] developed and validated an equation for estimating the GFR that can be used across the full age spectrum (FAS). The new FAS equation is

Table 1.2: Definitions of AKI, AKD, and CKD according to KDIGO criteria [111].

	Functional criteria	Structural criteria
AKI	Increase in SCr by 50% within 7 days, OR Increase in SCr by 0.3 mg/dl (26.5 mmol/l) within 2 days, OR Oliguria	No criteria
AKD	AKI, OR GFR < 60 ml/min per 1.73 m^2 for < 3 months, OR Decrease in GFR by $\geq$ 35% or increase in SCr by >50% for <3 months	Kidney damage for < 3 months
CKD	GFR <60 ml/min per 1.73 m^2 for >3 months	Kidney damage for >3 months

based on normalized serum creatinine (SCr/Q), where Q is the median SCr from healthy populations to account for age and sex. Coefficients for the equation are mathematically obtained by requiring continuity during the pediatric-adult and adult-elderly transition. Recently, Pottel et al. [148] developed and validated a modified FAS SCr-based equation combining design features of the FAS and CKD-EPI equations named the European Kidney Function Consortium (EKFC). According to an article published in 2021, the inclusion of race in equations to estimate GFR has become controversial. Alternative equations have been proposed to achieve similar accuracy without using race [79]. The current definition of AKI is flawed in that it is entirely based on an increase in SCr or a decrease in urine volume. Creatinine levels are often not indicative of GFR following injury since numerous renal and nonrenal factors influence creatinine levels. SCr is affected by some factors such as muscle mass, moreover, studies have shown that there are age and gender differences in creatinine generation [192]. Cystatin C (CysC) is another biomarker of kidney function that has gained the attention of researchers and clinicians over the past few years. In spite of the fact that it does not depend on muscle mass, the associated laboratory costs are ten times greater than those associated with creatinine testing. Therefore, a need to stratify patients according to what biomarker should be used to estimate their GFR at a given time. AKI contributes to adverse short-term and long-term outcomes. Different studies have linked AKI to the development of acute kidney disease (AKD), chronic kidney disease (CKD), end-stage kidney disease, longer hospitalization time, cardiovascular disease (CVD), and other complications, suggesting that even a short episode of acute kidney injury might lead to long term morbidity [178] and mortality [132],[172]. Table 1.2 shows the definitions of AKI, AKD, and CKD according to KDIGO [111].

Moreover, it has been reported that critically ill patients with dialysis-requiring AKI experience mortality rates above 50% [137]. The mortality rate of this

sudden kidney failure in ICU is approximately 30%–50% depending on the medical record of the patient and the stage of AKI [8, 62].

The purpose of this Ph.D. is to address the ML challenges associated with AKI. First, by applying statistical analysis and ML models, we investigate the association between AKI development and in-hospital mortality [133]. Second, we investigated the existing statistical or ML models that predict renal insufficiency after AKI episodes in ICU/hospital. Then we evaluated the differences between cystatin C- and creatinine-based estimated GFR in the follow-up of patients recovering from a stage 3 AKI in the ICU. Lastly, by developing machine learning models we predicted the time to death and the risk of chronic kidney disease for AKI survivors. The goal was to obtain a risk profile for every AKI survivor at discharge from the ICU, such that a personalized follow-up scheme can be proposed.

1.2 Development of clinical prediction models

Today, in the era of personalized medicine, predicting the presence of a disease (diagnosis) or an event in the progression of the disease (prognosis) has become increasingly important. It is necessary to have computer-interpretable data that is consistently recorded within the time frame for which a prediction is to be made, as well as reliable data that can be calculated. It is common for an intensive care unit to generate extensive amounts of data from a variety of devices for each patient.

1.2.1 Electronic health records

Many sectors have been transformed by the advent of computerization in the past few decades. The integration of electronic devices into ICUs is now a common practice. Although this results in a deluge of information, it simplifies data visualization and improves the overall user experience.

The electronic health record (EHR) is a digital record of a patient's health information that is stored electronically in a systematic manner [59]. A variety of information may be stored in an EHR, including demographic information, medical history, medication and allergy information, immunization status, laboratory results, radiology images, vital signs, and billing information [33].

It has been widely acknowledged for several decades that EHRs are fundamental to improving the quality of care [98]. Moreover, in light of the fact that patient records are shared via network-connected, enterprise-wide information systems or other information exchange systems, medical research can be accelerated with faster and easier access to patient data, resulting in the ability to perform

secondary use of the data even at a multicenter level, thereby providing benchmarking opportunities.

At the AZ Groeninge hospital (Kortrijk, Belgium), the EHR is implemented through different types of Patient Data Management System (PDMS) like Metavision 5, IMDSof, and Israel that stores continuous data on a minute-by-minute basis. Each patient generates an average of 6.3 megabytes of data each day. A total of 31 ICU beds at a saturation rate of 85% during the entire year results in approximately 250 GB of data being generated by the ICU of the AZ Groeninge. Large clinical trial databases have been created with the support of the hospital's IT department.

1.2.2 Data analytics in ICU

Data analytics is the process of discovering trends and drawing conclusions about the information contained in datasets. In general, big data refers to digital data that is generated in high volume and variety and is accumulated at high speed, resulting in datasets like the ones containing EHRs, which are too large to be processed by traditional data processing techniques [163].

There has been an increase in the use of big data analytics in a variety of fields, which include genome sequencing, drug discovery, and healthcare among others. More specifically in the healthcare domain, data analytics techniques can be utilized in order to aid in decision-making processes related to treating, diagnosing, and discharging patients from the intensive care unit [154].

Analyzing big data involves the use of techniques derived from computer science, such as machine learning. Machine learning is a field of study that examines how computers learn from data and the creation of algorithms that facilitate this process. As a type of data analytics, predictive analytics uses historical and current data to forecast activity, behavior, and trends. An important aspect of predictive analytics is predictive modeling, which is a mathematical process for predicting future events or outcomes based on patterns identified in a set of input data.

1.2.3 Clinical predictive modeling

The use of machine learning algorithms allows for identifying complex patterns within large and prosperous datasets. Clinical applications are particularly well suited to this facility [167]. A lot of machine learning techniques are used in many ICU applications such as length of stay prediction [78], mortality prediction [3], prediction of sepsis [136], the necessity of mechanical ventilation [12] and identifying patients with similar needs and trajectories [199], prediction

of complications in ICU, and processing and monitoring of vital signs in ICU patients.

The data contained in an EHR includes clinical information as well as the associated temporal information. Also, EHR data is irregularly sampled, as there are no regular time intervals between the patients' visits. It can consist of many features with periodic recording patterns. In spite of the limited number of unique conditions, medications, procedures, and measurements recorded at each visit, this may lead to an extremely large feature space (input) space for a prediction model. As a result, predictive modeling usually incorporates both static and dynamic clinical features according to the purpose. Figure 1.1 schematizes the extraction and representation of EHR data in preparation for the development of a predictive model.

It is often the time until an event occurs that is the outcome of interest in clinical studies. An example includes the time to death, the time until the patient leaves the hospital, the time until the recurrence of the tumor, etc. A time-to-event analysis is also known as survival analysis [37].

Machine learning communities have recently become increasingly interested in survival data for a variety of reasons. Firstly, physicians in the healthcare field are interested in accurate prognostics that will inform them regarding a patient's likelihood of adapting to medical treatment. Secondly, standard survival models encounter some challenges when it comes to real-world datasets since the presence of high-dimensional data is quite common, e.g., EHR data and gene expression data, and these traditional methods are not able to efficiently deal with such high-dimensional data. Moreover, they cannot easily capture nonlinear relationships between the covariates.

It is important to note that, although machine learning techniques have been used for decades in healthcare, in recent years, greater emphasis has been placed on the explanation of machine learning models [2]. Users' trust in machine learning models is essential, especially in healthcare, where merely providing traditional machine learning metrics such as the area under the receiver operating characteristic curve (AUROC), precision, and recall will not suffice for the prediction of the outcome. Interpretability of model predictions is an important consideration when implementing and utilizing them by clinical providers and other healthcare decision-makers. The concept of interpretable machine learning refers to the use of machine learning models that provide explanations for how certain predictions are made. This trend has led to applications utilizing models that can be more easily interpreted, such as decision trees (DTs) and random forests (RFs) [170, 161].

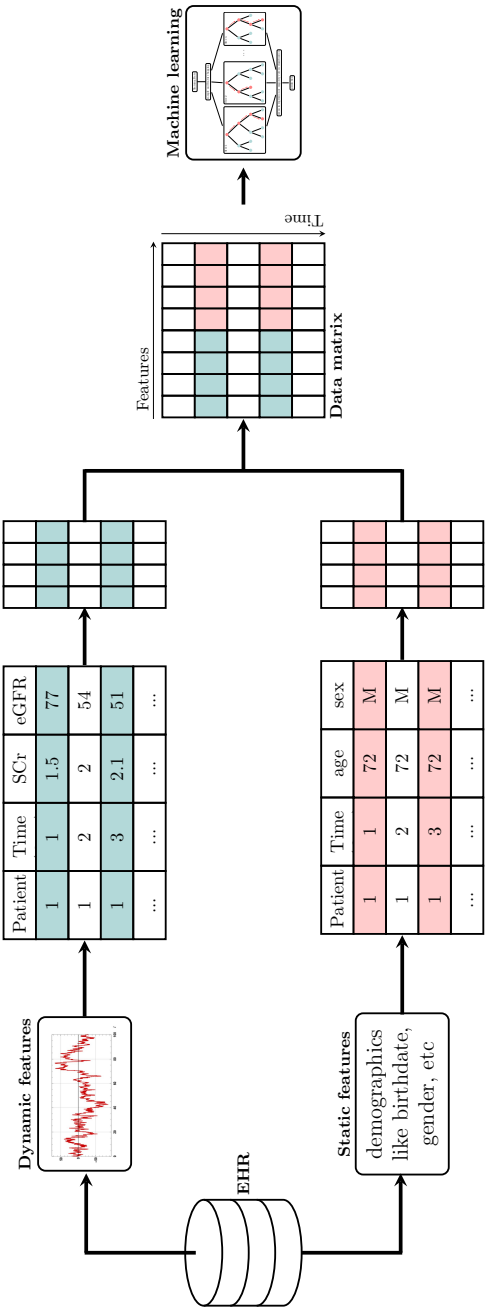


Figure 1.1: Overview of data extraction from EHR to develop a prediction model.

1.3 Thesis overview

The outline of the thesis is demonstrated in Figure 1.2. First, general background and review of some concepts of the employed models have been summarized in Chapter 3. Then, we present our goals and objectives in Chapter 2.

The core of our research is divided into three parts. The first part which includes Chapters 4, 5, and 6 is related to acute kidney injury including the effect of different definitions of AKI on mortality, the existing validated prediction models of AKI outcomes, and a comparison between the two most used kidney biomarkers and their effect on critically ill patients with AKI stage 3. The second part of this thesis focuses on the algorithmic design of novel machine learning models for time-to-event prediction. Our work is presented in Chapters 7 and 8. The last part of the conducted research was focused on predicting the outcomes of stage 3 AKI in critically ill patients after being discharged from the ICU. The outcome of the conducted work in this domain is presented in Chapter 9. Finally, a discussion of our work takes place in Chapter 10. More specifically, this thesis is organized as follows:

- In *Chapter* 2, the goals and objectives of our work are presented.
- In *Chapter* 3, several background concepts of the employed models are reviewed and discussed.
- *Chapter* 4 focuses on diagnosing acute kidney injury events based on different consensus definitions and their association with in-hospital mortality.
- In *Chapter* 5, existing validated risk prediction models for developing poor renal outcomes after AKI scenarios are presented and reviewed.
- *Chapter* 6 compares GFR estimation using SCr and CysC in detecting CKD over a 1-year follow-up after an AKI-stage 3 event in the ICU, as well as analyzes the association between eGFR (using SCr and CysC) and mortality after the AKI event.
- In *Chapter* 7, we present our new machine learning method for incorporating unlabeled data in a time-to-event analysis. This model, apart from fully observed and censored instances, also includes unlabeled instances. We propose three approaches to deal with this novel setting and provide an empirical comparison over fifteen real-life clinical and gene expression survival datasets.
- In *Chapter* 8, a new time-to-event prediction model is presented. We have transformed the time-to-event prediction problem into a semi-supervised regression problem in which we use a self-training wrapper approach with random survival forests as the base learner.

- *Chapter* 9 presents our ICU applications: outcome prediction of stage 3 AKI in critically ill patients after being discharged from the ICU. Specifically, we describe machine learning models for predicting CKD after 3 months and 6 months of developing AKI stage 3 in the intensive care unit. In addition, to estimate the mortality time for patients with AKI stage 3, a mortality time prediction model has been developed and validated on an external validation set.
- In *Chapter* 10, the main findings of this thesis are summarized and possible topics for future research are discussed.

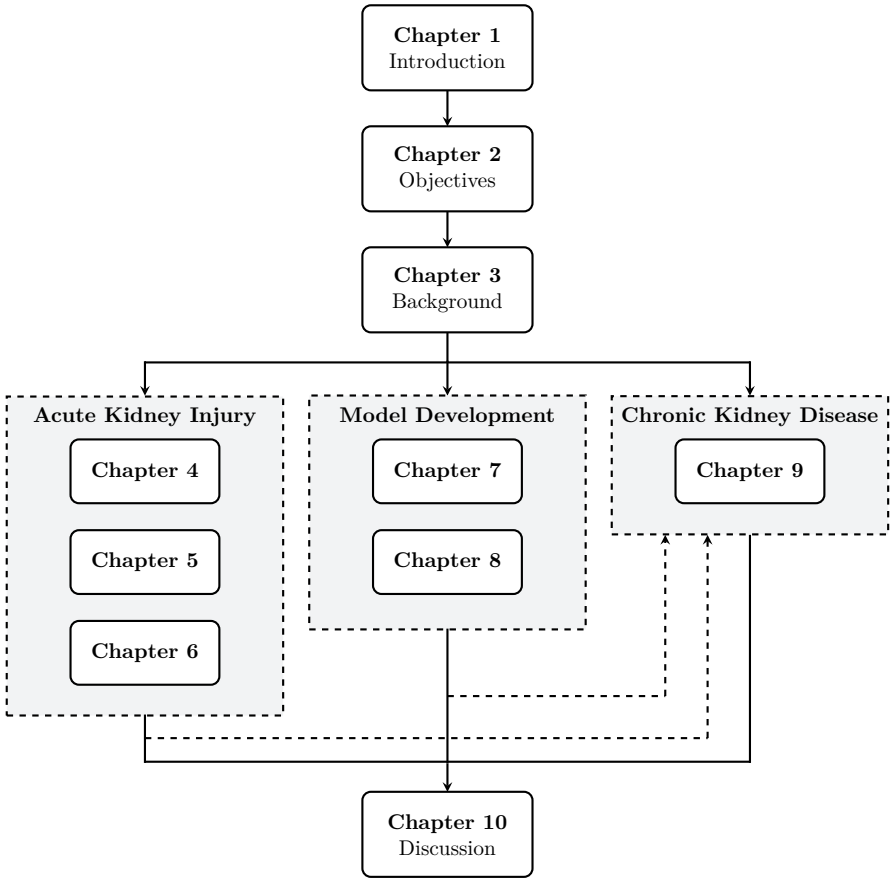


Figure 1.2: Thesis overview.

Chapter 2

Goals and Objectives

2.1 General objective

The general aim of this thesis is to develop new machine learning methods to address challenges presented in time-to-event analysis applications for critically ill patients with acute kidney injury in ICU. More specifically, we address the societal and economic challenge posed by ICU-related AKI by developing machine learning models to predict the risk of chronic kidney disease for AKI survivors. The goal is to obtain a risk profile for every AKI survivor at discharge from the ICU, such that a personalized follow-up scheme can be proposed. The fundamental contributions of this Ph.D. project include the adoption of existing and the development of new machine learning techniques. The project objective can be separated into methodological objectives, which make a substantial contribution to the field of machine learning, and medical application objectives, which may lead to an improved post-ICU policy for AKI patients.

2.2 Specific objectives

Our general objective is translated into specific objectives as follows:

Objective 1: Different definitions of diagnosing acute kidney injury and their association with in-hospital mortality.

The existence of different definitions of AKI makes it difficult to analyze the incidence and outcomes associated with AKI. Similar to AKIN guidelines, Kidney Disease: Improving Global Outcomes (KDIGO) clinical practice guidelines developed in 2012 categorize acute kidney injury (AKI) into three stages based

on the level of serum creatinine or the level of urine output. However, Sparrow et al. [177] investigated the impact of categorizing AKI stage 1 into two stages based on serum creatinine criteria and consequently modified the standard AKIN and KDIGO definitions. Consequently, the resulting 4-stage classification would affect the association of AKI stages with clinical outcomes Our first objective is to investigate the incidence of AKI events defined by 4 different definitions (standard AKIN and KDIGO, and modified AKIN-4 and KDIGO-4) and its association with in-hospital mortality.

Objective 2: Determine the optimal biomarker for estimating glomerular filtration rate.

The glomerular filtration rate (GFR) is a measure of how well the kidneys are working. The default estimation of GFR is done by inserting serum creatinine (SCr) values in a mathematical equation, where higher SCr values are associated with a decreased GFR function. However, ICU patients often suffer from muscle loss, e.g., as a result of long-term immobilization, which leads to a decrease in SCr. Therefore, the estimated GFR is unreliable. An alternative biomarker for estimating GFR is cystatin C (CysC). Unlike SCr, it is not affected by muscle mass; however, the associated laboratory costs are ten times higher. Thus, there is a need to stratify patients according to which biomarker should be used for them to estimate their GFR at a given time point. Our second objective is to compare GFR estimation using SCr and CysC in detecting CKD over a 1-year follow-up after an AKI-stage 3 in the ICU, as well as to analyze the association between eGFR (using SCr and CysC) and mortality after the AKI event.

Objective 3: Develop a machine learning-based model to predict time-to-event.

Surprisingly, survival data has not received much attention in data mining or machine learning communities. It is not possible to apply standard machine learning techniques directly to survival data due to censoring. Nevertheless, several studies transform the survival data into a format suitable for standard machine learning techniques, inevitably losing information. Often, the task becomes a binary classification task (does the patient survive a particular time point?), and censored data points are either removed, or their impact is decreased through a weighting technique. Several studies use semi-supervised learning techniques to deal with censored survival data. They treat the censored data points as unlabeled, thereby ignoring the survival information that they represent. Our third objective is to transform the time-to-event prediction problem into a semi-supervised regression problem in which censored observations are introduced as partially labeled observations since their target values should exceed the censoring time.

Objective 4: Develop machine learning methodology to include unlabeled data in time-to-event prediction.

It is challenging to follow up with patients after a hospital stay in many clinical studies because once patients resume their normal activities, it is often difficult to reach or motivate them to continue participating in the study. Consequently, dropouts are frequent, and follow-up data for many patients aren't available (only data from their hospital stay). In many of these prospective studies, the training set can easily be augmented with retrospective hospital data obtained from patients who were not enrolled in the study. If the study outcome is determined during follow-up, for both scenarios, this means that we often have a considerable unlabeled part of the training set (equivalently, the censoring time is zero for these patients). Based on successes in the semi-supervised learning domain, our goal is that unlabeled data, which is often easy to obtain, can increase the predictive performance in a survival prediction task. We propose three approaches to deal with this novel setting and provide an empirical comparison over real-life clinical and gene expression survival datasets.

Objective 5: Develop a risk predictor for the development of chronic kidney disease in AKI ICU patients.

Despite the high incidence rate of AKI among ICU patients and the considerable social and economic consequences, patients who are no longer dependent on dialysis often disappear from nephrologist follow-up once they leave the ICU (and hospital). Therefore, it is essential to accurately estimate the risk of progression for these patients, in order to develop a rational medical care policy and follow-up plan. Our fifth objective is to develop machine learning-based models to predict outcomes of acute kidney injury in critically ill patients.

Chapter 3

Background

3.1 Time-to-event analysis

Survival analysis is a widely used subfield of statistics that was originally designed to predict the lifespan of patients in a clinical setting. Additionally, survival analysis techniques may be applied more broadly to predict any time to event, such as the onset of a disease, failure times in an engineering context, etc. Survival analysis is primarily concerned with predicting a time-to-event distribution based on the presence of features, identifying factors that influence the distribution, and defining the nature of these factors.

3.1.1 Survival data and censoring

In survival analysis, during the study of the problem, omission of events of interest in some instances may happen. In some cases, this may be the result of a limited observation period or missing traces caused by uninteresting events that occurred elsewhere, or simply drop-out of the subjects from the study. Censoring refers to this concept, the most challenging aspect of survival data [100]. Censoring can generally be divided into three categories based on the reasons behind it: (i) *right-censoring*, where the observed survival time is less than or equal to the true survival time; (ii) *left-censoring*, where the observed survival time is greater than or equal to the true survival time; and (iii) *interval censoring*, where we only know that the event occurs during a given time interval [105, 202]. An example of censoring and survival data structure is shown in Figure 3.1. The figure illustrates that only subjects S2 and S5 experienced the event, and subject S4 is censored since there was no event during the

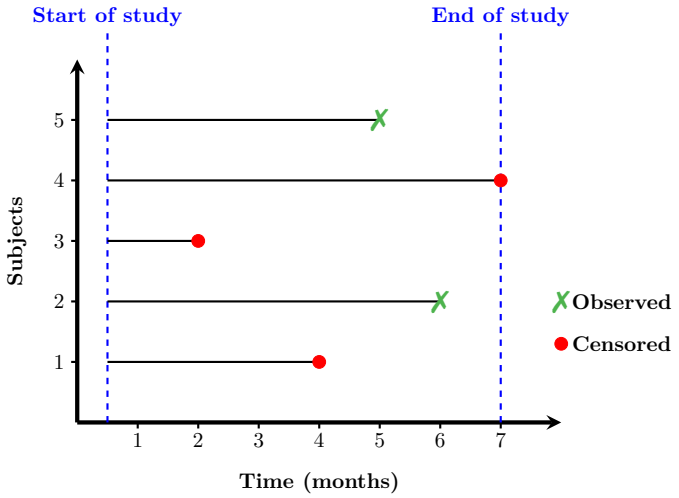


Figure 3.1: An illustration presenting the censoring problem in survival analysis.

study period, while subjects S1 and S3 are censored due to drop-outs or loss of follow-up during the study period.

Problem Statement: For a given instance i, represented by a triplet (X_i, y_i, δ_i), where $X_i \in \mathbb{R}^P$ is the feature vector; δ_i is the binary event indicator (i.e., $\delta_i = 1$ for an uncensored instance and $\delta_i = 0$ for a censored instance); and y_i denotes the observed time and is equal to the survival time T_i for an uncensored instance and C_i for a censored instance; that is,

$$y_i = \begin{cases} T_i, & \text{if } \delta_i = 1. \\ C_i, & \text{if } \delta_i = 0. \end{cases} \tag{3.1}$$

In survival analysis, the objective is to estimate the time to the event of interest T_j for a new instance j based on feature predictors described by X_j. It should be noted that in survival analysis problems, T_j and C_i are both non-negative and continuous [202].

3.1.2 Survival and hazard function

In survival analysis, one of the primary goals is to determine the survival function. The survival function is used to represent the probability that the time to the event of interest is not earlier than a specified time t [105, 100].

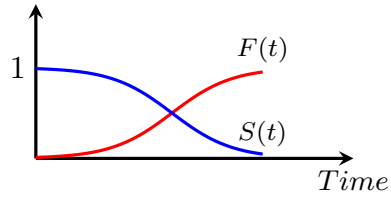


Figure 3.2: Cumulative distribution function $F(t)$ and survival function $S(t)$.

$$S(t) = P(T \geq t). \tag{3.2}$$

Alternatively, the cumulative death distribution function $F(t)$, which represents the probability of the event occurring before t, is defined as $F(t) = 1 - S(t)$. An illustration showing the relationships between these functions is shown in Figure 3.2.

The hazard function $h(t)$, sometimes also known as the instantaneous death rate is defined as $h(t) = f(t)/S(t)$, where $f(t)$ is the density function for the time to an event and $f(t) = -\frac{d}{dt}S(t)$. More specifically, $h(t)$ represents the likelihood of the event occurring at time t given that no event has occurred before time t [44]. Similar to $S(t)$, $h(t)$ is also a non-negative function. The Cumulative Hazard Function (CHF) is defined as $H(t) = \int_0^t h(u)du$ which results in the following equation:

$$S(t) = exp(-H(t)). \tag{3.3}$$

3.1.3 Cox regression

A survival/hazard function can be estimated using three different types of statistical methods: non-parametric, semi-parametric, and parametric. In the semi-parametric category, the Cox model [34] is the most commonly used regression analysis approach for survival data. In spite of being based on a parametric regression model, the Cox model is described as semi-parametric due to the fact that no knowledge of the underlying distribution of time to the event of interest is required [202]. The survival function in a Cox model is computed as follows:

$$S(t) = S_0(t)^{exp(X\beta)}, \tag{3.4}$$

where $S_0(t)$ represents the baseline survival function, $X_i = (x_{i1}, x_{i2}, ..., x_{iP})$ is the corresponding feature vector for instance i; and $\beta^T = (\beta_1, \beta_2, ..., \beta_P)$ is the coefficient vector.

To estimate these coefficients, a partial likelihood is defined as follows:

$$L(\beta) = \prod_{j=1}^{N} [\frac{exp(X_j\beta)}{\sum_{i \in R_j} exp(X_j\beta)}]^{\delta_j}, \tag{3.5}$$

where N is the number of instances.

The development of data collection and detection techniques have led to the accumulation of high-dimensional data in many real-world domains. There may be datasets in which the number of features (P) in a given set of data is almost equal to, or may even exceed, the number of instances (N). Therefore, it is difficult to construct a good prediction model that incorporates all the available information in the feature set. In this regard, several different penalty functions including the Lasso (least absolute shrinkage and selection operator) [187], the Ridge [70], and the Elastic-Net (EN) [214] have been developed for the purpose of identifying the features that are most relevant to the outcome variable among what can be tens of thousands of features.

In Lasso, features are selected and regression coefficients are estimated simultaneously, using l_1-norm regularization. Lasso-Cox regression inherits the properties of the l_1-norm in feature selection for both fitting and penalization of the coefficients.

Ridge-Cox regression incorporates an l_2-norm regularizer to select correlated features and shrink their values toward each other.

EN-Cox regression [175] uses the EN properties which combine the l_1 and squared l_2 penalties and has the potential to perform feature selection and deal with the correlation between the features simultaneously. In the context of survival analysis, we have compared our proposed methods with COX regression. The Weibull model and the accelerated failure time model are also possible alternatives, but in this thesis, the COX model is used as our baseline model since the Cox regression model is applicable to a wider class of distributions and it is a semi-parametric model while the Weibull regression model is fully parametric.

3.2 Machine learning

Finding patterns in data is a fundamental problem that has been studied and solved extensively for a long period of time. Computer algorithms have been used

to find regularities in data and to use these regularities for taking actions, such as classifying the data. The field of machine learning focuses on the automatic discovery of regularities in data through the use of algorithmic processes [13]. According to Tom Mitchell [127], machine learning can be described as follows: "A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

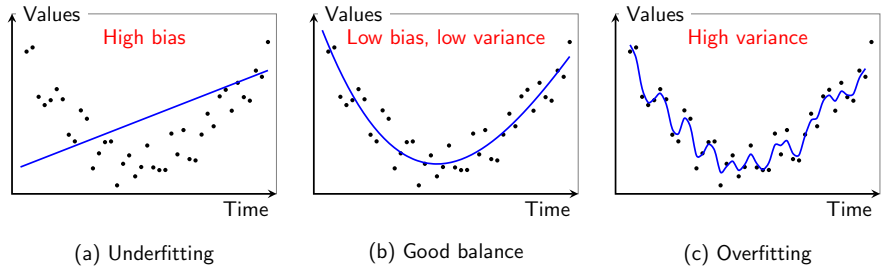
The field of machine learning can be divided into three main categories: supervised learning, unsupervised learning, and semi-supervised learning. The most common form of machine learning is supervised learning [131]. In this problem, the task T is to learn a mapping f from inputs $x \in X$ to outputs $y \in Y$. More precisely, the instances (e.g., patients, genes, etc) are described by input variables and are also associated with one or more output variables. The inputs x are also called the features, covariates, or predictors that are often fixed-dimensional vectors of numbers, such as the height and weight of a person, or the pixels in an image. In this case, $X = \mathbb{R}^D$, where D is the dimensionality of the vector (i.e., the number of input features). The output y is also known as the label, target, or response. The experience E is given in the form of a set of N input-output pairs $D = (x_n, y_n)_{n=1}^N$ known as the training set and N is called the sample size [131].

Predictive tasks are defined by the type of Y. When the output is numerical, such as the price of an object, the predictive task is known as regression. Classification refers to tasks that produce categorical results, such as classifying types of flowers.

In supervised learning, the goal is to automatically develop classification models that learn a function ($f : X \to Y$) that accurately predicts the labels for given data input [208]. In order to achieve optimal performance, most models require tuning of a number of parameters. The model will be inaccurate if it ignores the regularities of the training data. On the other hand, if the model is too complex, it will capture all the regularities in the training data, including noise and randomness. Therefore, the generated model will not be able to generalize effectively to new data. *Underfitting* is the first case and *overfitting* is the second.

Before we proceed, it is essential to explain two terms:

Bias: An assumption that is made by a model to facilitate the learning of a function. In other words, it is the error rate associated with the training data. When the error rate is high, we refer to it as high bias, and when it is low, we refer to it as low bias.



(a) Underfitting (b) Good balance (c) Overfitting

Figure 3.3: Illustration of a supervised learning problem: (a) an underfitted model, (b) a well-fitted model, (c) an overfitted model.

Variance: Variance is the difference in error rates between training data and testing data. If there is a large difference between the errors, it is referred to as a high variance, whereas if there is a small difference, it is referred to as a low variance. A low variance is usually desirable for generalized models.

An underfitted model is one that performs poorly on training data and is unable to generalize to new data, while an overfitted model is one that performs perfectly on training data and does not generalize to new data. Figure 3.3 is an example of a regression problem. It demonstrates the problems of underfitting, well-fitting, and overfitting and how we can use linear regression with polynomial features to approximate nonlinear functions. In Figure (a), we can see that a linear function (polynomial with degree 1) is unable to capture the patterns in the data; however, higher degrees of the model will overfit the training dataset, i.e. it learns the noise of the training data (figure (c)).

Besides supervised learning, in other machine learning problems, the training data consists of a set of input vectors x without any corresponding target values. A common goal of such unsupervised learning problems is the discovery of clusters of similar instances within the data. This is called clustering [13].

Finally, the last category of machine learning problems, which is called semi-supervised learning, falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data). During training, semi-supervised learning combines a small amount of labeled data with a large amount of unlabeled data [194]. Semi-supervised learning is applied in a variety of fields from healthcare, education, entertainment, etc. Particularly in healthcare, where collecting labeled data is challenging, semi-supervised learning methods are pretty common and useful. In this thesis, we have addressed these challenges using semi-supervised learning, as further specified in Chapters 7 and 8.

3.3 Machine learning models

It is well known that there are a wide variety of machine learning models in the literature, ranging from biologically inspired models, such as deep learning [103], to statistical models, such as logistic regression, support vector machines (SVMs) [32], k-Nearest Neighbor [64]. Despite being very different from one another, all methods present a trade-off between interpretability and performance.

The logistic regression method is the most commonly used baseline method for binary classification problems (problems where there are two possible values). The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1:

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + ... + \beta_P x_P^{(i)})}, \tag{3.6}$$

where $X_i = (x_1^{(i)}, x_2^{(i)}, ..., x_P^{(i)})$ is the corresponding feature vector for instance i; and $\beta^T = (\beta_0, \beta_1, ..., \beta_P)$ is the coefficient vector.

As a baseline model for classification, the logistic regression method is utilized in this thesis.

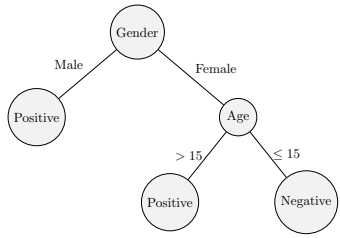
Models based on decision trees are the subject of this thesis. Several reasons contribute to the wide use of tree-based learners, including their non-parametric nature, ability to capture complex non-linear relationships and interactions, and ease of interpretation. Specifically, model interpretability and explainability are crucial for clinical and healthcare practice since in general, the main aim of these interpretability techniques is to shed light and provide insights into the prediction process of the machine learning models and to be able to explain how the results from the prediction were generated [162].

3.3.1 Decision Trees

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks [18, 17]. In order to construct decision trees (DTs), data are typically recursively partitioned, and then several metrics can be used to decide the best feature split in a top-down greedy approach[18, 14]. A decision tree consists of nodes connected by edges. Each node has an incoming edge that connects it to its parent node and an outgoing edge that connects it to its children. In a tree, the root is the top node and the leaves are the nodes that have no output edges. An example of a decision tree and its corresponding decision boundary is demonstrated in Figure 3.4.

Instance	Age	Sex	Label
1	15	Female	Negative (0)
2	22	Male	Positive (1)
3	12	Female	Negative (0)
4	21	Male	Positive (1)

(a)



(b)

Figure 3.4: An example of a decision tree. In Figure (b), we present an example of a decision tree generated using the dataset from Figure (a).

3.3.2 Tree ensembles

Ensemble methods in statistics and machine learning learn and combine multiple models to provide better prediction performance than could be obtained by any individual model alone [141].

Random forests (RFs) [17] and random survival forests (RSFs) [84] are widespread ensemble learning methods that work based on a collection of multiple decision trees, as displayed in Figure 3.5.

Random forest

Random forest uses the bagging principle. The bagging process selects a random sample from the data set. A model is therefore generated by replacing the samples (bootstrap samples) provided by the original data with a replacement methodology called row sampling. The process of row sampling with replacement is known as bootstrapping. A decision tree is trained on each bootstrap sample independently. Every node of every tree is split by computing the best possible split among a random subset of selected feature candidates. The final output is based on majority voting after combining the results of all models for the classification task and averaging over each tree output for the regression task. The step that involves combining all the results and generating output is known as aggregation [17].

Random survival forest

A random survival forest (RSF) is an extension of the random forest paradigm to censored survival data [84]. It is based on the random forest approach but modifies the split criterion and the prediction at each leaf node to accommodate censored survival data.

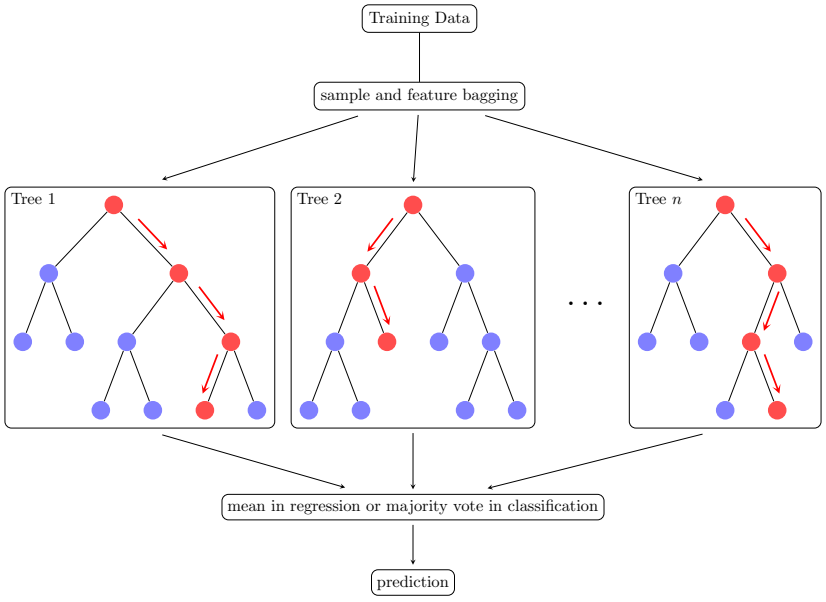


Figure 3.5: An illustration of a tree-ensemble. The dataset is split into several subsets and one tree is built with each subset. The final prediction is given by combining the predictions of all trees.

In the same way as random forests, RSF combines bootstrapping, tree building, and prediction aggregation. RSF, however, explicitly considers survival time and censoring information in its splitting criterion for growing trees. The RSF consists of three main steps. Initially, it generates B bootstrap samples from the original data. After the bootstrap sample has been determined, a survival tree is grown for each sample. A tree is constructed by randomly selecting p candidate variables at each node, where p is a feature set dimension. The task is to split the node into two child nodes using the best candidate variable and split point, as determined by the log-rank test [166]. The best split is the one that maximizes survival differences between the two child nodes. Growing the obtained tree structure is continued until a stop criterion holds (e.g., until the number of observed instances in the terminal nodes drops below a specified value). In the last step, the cumulative hazard function (CHF) associated with each terminal node in a tree is calculated by the Nelson-Aalen estimator, which is a non-parametric estimator of the CHF [89]. All cases within the same terminal node have the same CHF. The ensemble CHF is constructed as the average over the CHF of the B survival trees.

3.4 Semi-supervised learning

In semi-supervised learning algorithms, unlabeled examples are used in addition to labeled ones in order to develop models that perform better than those that use only labeled examples. A simple approach to extending existing, supervised algorithms to the semi-supervised setting is to first train on labeled data, and then use the predictions of the resulting model to generate additional labeled data. The model can then be re-trained on this pseudo-labeled data in addition to the existing labeled data. Such methods are known as wrapper methods. A significant advantage of wrapper methods is that they can be used with almost any supervised base learner. One of the most basic methods of pseudo-labeling is self-training [190]. There also exist a number of semi-supervised algorithms that use decision trees to exploit their desirable properties [184, 107].

3.5 Evaluation metrics

An integral part of any machine learning project is the evaluation of the algorithm after the model has been built. We can determine the generalization ability of a machine learning model based on the metrics used to evaluate it. It is also essential to choose the right metric when evaluating machine learning models. In different applications, machine learning models can be evaluated using a variety of metrics. In this thesis, we have used classification and time-to-event prediction models; hence we introduce the metrics related to these categories.

3.5.1 Classification metrics

The metrics that were most often used in this thesis for classification models are accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUROC) [47]. In their simplest case, these measures are applied in a binary classification where instances may be positive (class 1) or negative (class 0). More specifically, a prediction may be defined as true positive (TP), false positive (FP), true negative (TN), and false negative (FN). When the model correctly predicts the label associated with an instance, TPs and TNs are obtained. A TP is the number of accurate predictions of a positive class, while a TN is the number of accurate predictions of a negative class. FPs, on the other hand, are inaccurate predictions regarding instances whose classes are negative, but which were predicted as positive by the model. In the same way, FN represents the opposite case when the expected class is positive, but the model classifies it as negative. Accuracy, precision, and recall are further measured using these values.

Accuracy is a ratio of correctly predicted observations to the total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.7}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP}. \tag{3.8}$$

Recall or sensitivity is the ratio of correctly predicted positive observations of all observations in the actual class.

$$Sensitivity = Recall = \frac{TP}{TP + FN}. \tag{3.9}$$

Specificity is the ratio of correctly predicted negative observations of all observations in the actual class.

$$Specificity = \frac{TN}{FP + TN}. \tag{3.10}$$

F1-Score is the harmonic average of Precision and Recall.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{3.11}$$

A predictive model often outputs probabilities of an instance belonging to a specific class, for example, a 75 percent probability that a patient will develop a particular disease. To obtain a final prediction, a threshold must be applied. Typically, 50% is selected, resulting in instances with probabilities below 50% being classified as negative, and positive in the opposite case.

The ROC curve is defined as the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. TPR is equal to sensitivity as defined in Equation 3.9. FPR is equal to 1- Specificity with Specificity defined in Equation 3.10. The ratio represents the amount of incorrectly classified negative instances in relation to the total number of negative instances.

A precision-recall curve (or PR Curve) is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds. An algorithm should have both a high level of precision and a high level of recall. It should be noted, however, that most machine learning algorithms involve a trade-off between the two. In general, a good PR curve has a higher AUC (area under the curve).

(a) (b)

Figure 3.6: Order graphs representing the ranking constraints. (a) No censored data and (b) with censored data. The empty circle represents a censored point. The points are arranged in the decreasing value of their survival times with the lowest being at the bottom.

3.5.2 Survival analysis metrics

The metric that was most often used in this thesis for time-to-event models is the Concordance index (CI) which is the most commonly used metric for evaluating survival models and represents the generalization of the ROC curve over all data in the survival analysis [61]. In survival analysis, instead of measuring the absolute survival time for each instance, a popular way to assess a model is to estimate the relative risk of an event occurring for different instances. The Harrell's concordance index (C-index) [61] is a common way to evaluate a model in survival analysis [165]. C-index can be interpreted as the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered. In other words, it is the probability of concordance between the predicted and the observed survival time. As shown in figure 3.6, two subjects' survival times can be ordered not only if (1) both of them are observed but also if (2) the observed time of one is smaller than the censored survival time of the other [180]. Consider a set of observation and prediction values for two different instances, $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$, where y_i and $\hat{y}_i$ represent the actual survival time and the predicted value, respectively. The concordance probability between these two instances can be computed as:

$$Concordanceindex = Pr(\hat{y}_1 > \hat{y}_2 | y_1 > y_2). \tag{3.12}$$

3.6 Evaluation strategies

A machine learning model must always be validated for performance evaluation. Basically, data validation involves determining whether the results quantifying hypothesized relationships between the features can be used to interpret the data. In general, an evaluation metric is performed after the model has been

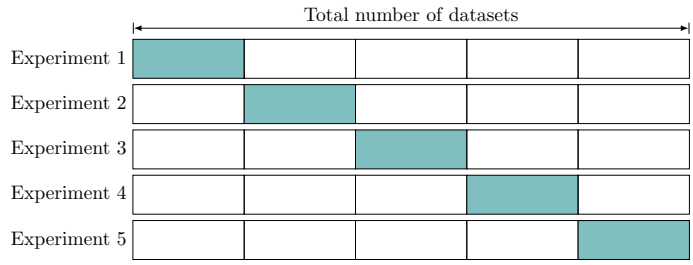


Figure 3.7: An example of 5-fold cross-validation where the testing folds are highlighted in green, whereas the training ones are represented in white. The the overall performance consists of the average performance obtained in the 5 folds.

trained, a process known as residual evaluation. As a result of this process, a numerical estimate is made of the difference between predicted and original responses, also known as the training error. It is important to note, however, that this is only an indication of how well our model performs on data that was used to train it. The data may be underfitted or overfitted by the model. Consequently, this evaluation technique lacks an indication of how well the learner will generalize their knowledge to unknown or independent data. It is possible to resolve this problem by removing a portion of the training data and using it as a test for the model trained on the remaining training data. This is known as the holdout method. In the absence of sufficient data, underfitting can occur when a portion of the data is removed for validation. As a result, we are in need of a method that will provide sample data for the training of the model as well as enough data for validation. Many methods are presented in the literature for splitting the data, including group k-fold, time series split, leave-one-out cross-validation, and k-fold cross-validation. The most highly preferred strategy is the k-fold cross-validation [53], which is illustrated in Figure 3.7.

When performing k-fold cross-validation (k-fold CV), the dataset is divided randomly into k equal subsets. Following that, a subset will be selected as the test subset for evaluating the predictions. A model is trained using the remaining k-1 subsets. This procedure is repeated k times, using a different subset as the test subset each time. It is ultimately determined by averaging the k obtained results from each fold to obtain the final evaluation result.

Cross-validation may produce a false impression of performance when the dataset is imbalanced, as certain outputs may be underrepresented in partitions, leading to an inaccurate estimation of the performance. Stratified cross-validation is preferable in these situations. It is similar to regular cross-validation in that stratified cross-validation divides the data into multiple folds, but relative label

frequencies are maintained across folds regardless of the type of analysis [50].

As a side note, cross-validation is not always possible when a limited amount of data is available, leading to partitions that are not sufficient for the construction of predictive models. A common cross-validation method in this situation is leave-one-out cross-validation. During leave-one-out cross-validation, K is fixed to the number of instances, which results in the testing fold consisting of one instance, while the training fold is constructed from the remaining instances [130]. Although leave-one-out cross-validation provides better reliability in this situation, when applied to large datasets, it is associated with high computational costs.

Frequently, cross-validation is used to assess the effectiveness of a machine learning model, particularly when it is necessary to mitigate overfitting. Additionally, it facilitates the determination of the hyperparameters of the model, as it allows us to identify which parameters will result in the lowest test error.

Chapter 4

The effect of different consensus definitions on diagnosing acute kidney injury events and their association with in-hospital mortality

The following chapter has been published in the Journal of Nephrology:

Nateghi Haredasht, F., Antonatou, M., Cavalier, E., Delanaye, P., Pottel, H. & Makris, K.The effect of different consensus definitions on diagnosing acute kidney injury events and their association with in-hospital mortality. *Journal of Nephrology.* 2022 Apr 20:1-9.

DOI: 10.1007/s40620-022-01323-y.

Abstract

Backgrounds: Due to the existence of different AKI definitions, analyzing AKI incidence and associated outcomes is challenging. We investigated the incidence of AKI events defined by 4 different definitions (standard AKIN and KDIGO, and modified AKIN-4 and KDIGO-4) and its association with in-hospital mortality.

Methods: A total of 7242 adult Greek subjects were investigated. To find the association between AKI stages and in-hospital mortality, we considered both the number of AKI events and the most severe stage of AKI reached by a patient adjusted for age, sex, and AKI staging using multivariable logistic regression. To predict mortality in AKI patients defined by the four definitions, a classification task with two prediction models (random forest and logistic regression) was also conducted.

Results: The incidence of AKI using the KDIGO-4 was 6.72% for stage 1a, 15.71% for stage 1b, 8.06% for stage 2, and 2.97% for stage 3; however, these percentages for AKIN-4 were 11%, 5.83%,1.75%, and 0.33% for stage 1a, stage 1b, stage 2, and stage 3, respectively. Results showed that KDIGO-4 is more sensitive in detecting AKI events. In-hospital mortality increased as the stage of AKI events increased for both KDIGO-4 and AKIN-4; however, KDIGO-4 (KDIGO) has a higher odds ratio at a higher stage of AKI compared to AKIN-4 (AKIN). Lastly, when using KDIGO, random forest and logistic regression models are performing almost equally with a c-statistic of 0.825 and 0.854, respectively.

Conclusion: The present study confirms that within the KDIGO AKI stage 1, there are two subpopulations with different severity of clinical outcomes (mortality).

4.1 Introduction

Acute Kidney Injury (AKI) is a heterogeneous clinical syndrome associated with various clinical presentations and characterized by a rapid deterioration of kidney function [16, 120, 21, 75, 159]. This syndrome is associated with considerable morbidity, mortality, and high healthcare costs [155]. AKI may also lead to the development of chronic kidney disease (CKD) or end-stage renal disease (ESRD) [164, 26, 24]. The incidence of AKI is increasing worldwide, particularly among hospitalized patients with acute illness and those undergoing major surgery [173, 142, 174, 95]. The main causes for this increase could be attributed to the increase in the number of patients hospitalized who are susceptible to this disease: [aging population, increased incidence of cardiovascular disease, diabetes mellitus, and CKD], and to an expanding characterization of modifiable risk factors, such as sepsis, administration of contrast media and exposure to nephrotoxins and nephrotoxic medications [120, 10, 138]. However, the incidence of AKI varies widely in reported studies, which likely reflects differences in case ascertainment, and the location of patient care, but the choice of the definition retained for AKI could also impact this incidence [212, 57, 179, 171, 176]. Since 2002 when the first consensus criteria of AKI (known as Risk, Injury, Failure, Loss-of-kidney-function, and End-stage kidney disease or RIFLE) were proposed, a major step has been made toward a uniform diagnostic approach to AKI [11]. This definition required a pre-morbid serum creatinine value which was lacking in many patients admitted with acute illness [211]. To address the limitations of RIFLE criteria, the Acute Kidney Injury Network suggested a modified definition, which focused on dynamic changes of SCr, more than on estimated GFR by equations, within a period of 48 hours at any time during a patient's hospitalization [124]. In order to calculate absolute and relative increases in SCr within a period of 48 hours the lowest SCr value during this period was used as the baseline for the calculations. In 2012 the Kidney Disease Improving Global Outcomes (KDIGO) published a clinical guideline with the aim to harmonize AKIN and RIFLE diagnostic criteria into one universal diagnostic guideline [92]. The new criteria combined the absolute increase in SCr of 0.3 mg/dL over a 48-hour period from the AKIN definition with the 50% relative increase in SCr over 7 days from the RIFLE definition into one set of criteria for AKI diagnosis. KDIGO also accepts the 3 stages model proposed by AKIN to categorize the severity of AKI. These combined diagnostic criteria in the KDIGO definition mean that the absolute increase in SCr over 48 hours and the relative increase over 7 days are equivalent criteria. However, several studies have questioned this equivalence as the relative increase criterion may overestimate the AKI diagnosis when the SCr baseline of the patient is low and the absolute criterion may underestimate AKI and vice versa [212, 177, 186, 121]. Some even suggested that the use of relative criteria to diagnose AKI might be inappropriate in patients with low baseline SCr [212]. Recently Sparrow et al., evaluated the

Table 4.1: KDIGO-4 and AKIN-4 definition of AKI [177].

	AKIN-4	KDIGO-4
Stage 1a	$\geq$ 0.3 absolute SCr increase over a 48-hour window of observation	$\geq$ 0.3 absolute SCr increase over a 48-hour window of observation
Stage 1b	$\geq$ 50% relative SCr increase over a 48-hour window of observation	$\geq$ 50% relative SCr increase over a 7-day window of observation
Stage 2	$\geq$ 100% relative SCr increase over a 48-hour window of observation	$\geq$ 100% relative SCr increase over a 7-day window of observation
Stage 3	$\geq$ 200% relative SCr increase over a 48-hour window of observation	$\geq$ 200% relative SCr increase over a 7-day window of observation

impact of further subcategorizing the KDIGO-defined AKI stage 1 into two stages based on SCr criteria: stage 1a (an absolute increase of SCr of 0.3 mg/dL within 48 hours) and stage 1b (a 50% relative increase in SCr within 7 days) and therefore creating a 4-stage KDIGO classification which they named as KDIGO-4 (see Table 4.1). A similar analysis was carried out using the same modification for the AKIN criteria. The present study aimed to investigate the incidence of AKI events defined by 4 different definitions (standard AKIN and KDIGO, and modified AKIN-4 and KDIGO-4) and its association with in-hospital mortality.

4.2 Materials and methods

4.2.1 Study design-Patient population

This study is a retrospective observational study where we used existing medical records data. The study was approved by the hospital's Ethical and scientific committee. All patients admitted to KAT General Hospital of Attiki in Athens, Greece, from January 1, 2016, to June 30, 2019, were screened for inclusion. Exclusion criteria included Age $<$ 18 years, patients with fewer than five SCr measurements during hospitalization, and hospital stay less than seven days. The time between admission and discharge was recorded as the actual hospitalization period. Any observation not lying within this period was discarded from the data. Patients with multiple admissions-discharges were included and were considered as separate cases. The hospital is a major trauma center, so a nephrology clinic or gynecology clinic does not exist. As a result, no pregnant woman neither a CKD patient stage 5 nor nephrectomized or kidney transplanted patients are admitted to this hospital. Hence, such patients are not included in the dataset. Finally, of 11382 hospital admissions, 7242 patients were included in this study (Figure A.1 in Appendix A).

4.2.2 Definitions - Acute Kidney Injury criteria and calculations

Only the creatinine-based criteria were considered because urine output was not available in all patients according to the AKIN score. AKI diagnosis can be made by either an absolute increase of 0.3 mg/dL (26 $\mu mol/L$) in SCr within 48 hours or a 50% increase from baseline again within the same timeframe. On the other hand, for the KDIGO criteria, the window of observations for the 50% increase from baseline is established over 7 days. In this study, the minimum SCr value within a rolling 48-hour window for each inpatient SCr value was defined as a dynamic baseline surrogate [204].

Staging of AKI is common in both definitions and three severity stages are defined in both definitions. According to AKIN criteria, stage 2 is defined as an increase of $\geq$ 2-3 times from baseline, and an increase in SCr up to 3 times from baseline is classified as stage 3. On the other hand, KDIGO defines stage 2 as an increase in SCr of $\geq$ 2-3 times and stage 3 up to 3 times from baseline within 7 days [120]. The primary focus of our study was to evaluate the equivalence of the absolute increase of 0.3 mg/dL (stage 1a) with the relative increase of 50% (stage 1b) in the KDIGO and AKIN criteria.

4.2.3 Outcomes: incidence of AKI, association with mortality, and mortality prediction

The primary outcome was to estimate the incidence of AKI events in our cohort and to evaluate the revision of KDIGO criteria into 4 stages as proposed by Sparrow et al [177]. We evaluated if there was any association between the number of AKI events and mortality, as well as the association between different stages of AKI and mortality. We also tested the effect of the revised 4 stages criteria on the association with the selected clinical outcomes. As our secondary objective, we applied a machine-learning algorithm to predict mortality in AKI patients. For this purpose, we employed a random forest model [17]. The results of the random forest model were compared with the logistic regression model.

4.2.4 Statistical analyses

Descriptive statistics for the AKI incidence and in-hospital mortality, based on the different definitions AKIN, KDIGO, AKIN-4, and KDIGO-4 are used and presented as percentages. A comparison of percentages is done with the chi-square test. To analyze the association between AKI events (as defined by the 4 different definitions) and mortality, we have considered two different approaches. The first approach was to consider the number of AKI events/stages for each patient. The second approach was to consider only one AKI episode (the most severe). For the first approach, we used variables age, gender, and the number of AKI events that patients had experienced. The model for the

second approach consisted of variables age, gender, and the different stages of AKI. We also cast a prediction task (classification) by classifying AKI patients based on the clinical outcome (mortality) using multivariable logistic regression and a random forest algorithm. In supervised learning, it is common to use at least two different models based on different mathematics, to confirm (or contra-indicate) the results (and the interpretation concerning the AKI events definitions). Moreover, it will aid researchers in the selection of an appropriate supervised machine learning algorithm for their studies.

Random forest

Random forest is an ensemble-based learning algorithm introduced by Breiman in 2001 [17]. The ensemble technique used by random forests is called Bagging (also known as Bootstrap aggregating). Figure A.2 in Appendix A illustrates an example of an ensemble decision tree.

4.3 Results

4.3.1 Patients

Of the 11382 hospital admissions with at least five SCr measurements, 7242 were included in the study after the exclusion of the patients who were under the age of 18 (n = 438) and had less than 7 days of hospitalization (n = 3702). Characteristics of patients are shown in Table 4.2. It is worth mentioning that the inclusion criteria were chosen able to fulfill the AKI criteria. Out of 7242 patients, 55% were females and the median (IQR) age of the cohort was 77 (18–102) years. The median length of stay was 16 (1–1171) days, and the mortality rate in the hospital was 9.5% (n= 689 patients). Moreover, the distribution of the GFR according to the CKD stages is presented in Table 4.2.

4.3.2 AKI incidence

Patients had a mean age of 72±17.4 years (range, 18 – 102). Forty-five percent of patients were male. The incidence of in-hospital AKI using KDIGO-4 was 6.72% for stage 1a, 15.71% for stage 1b, 8.06% for stage 2, and 2.97% for stage 3 (Table 4.3). Percentages for AKIN-4 were 11.5%, 5.83%,1.75%, and 0.33% for stage 1a, stage 1b, stage 2, and stage 3, respectively. The incidence of in-hospital mortality is also shown for both KDIGO and AKIN definitions in Table 4.3.

Note that patients may experience multiple AKI events during their hospital stay, with different grades of severity. Consequently, the number of events is not necessarily adding to the total of 7242, as the AKI events are counted. Figure

Table 4.2: Patient characteristics.

Characteristics	All patients (N=7242)
Female sex, n (%)	3986 (55.04%)
Median Age, years (IQR)	77 (18-102)
Median Length of stay (days), (IQR)	16 (1- 1171)
Admission department:	
Orthopedic clinic	4076 (56.28%)
ICU	1096 (15.13%)
General surgery	1008 (13.92%)
Cardiology	605 (11.6%)
other departments	457 (6.31%)
Median Creatinine (mg/dL) at admission:	
Females	0.80 (0.26-9.95)
Males	0.93 (0.38-13.84)
EKFC-eGFR (mL/min/1.73m^2) at admission:	
>90 (CKD1)	1214 (16.76%)
60-89 (CKD2)	3103 (42.85%)
45-59 (CKD3A)	1271 (17.55%)
30-44 (CKD3B)	1006 (13.89%)
15-29 (CKD4)	531 (7.33%)
<15 (CKD5)	117 (1.6%)

Values are median (IQR) or n(%).
ICU: intensive care unit; eGFR: estimated glomerular filtration rate; EKFC: European Kidney Function Consortium [148]; CKD: chronic kidney disease.

A.3 in Appendix A illustrates the presence of multiple events with various grades of severity for two random patients.

Table 4.3 also shows that 5713 out of 7242 patients did not experience an AKI event according to KDIGO-4, while this number was 6172 out of 7242 according to AKIN-4. Actually, 461 patients had no AKI events according to AKIN-4 but had AKI events according to KDIGO-4 (stage 1a was absent, but there was 634 stage 1b events, 156 stage 2, and 47 stage 3 events in these 461 patients), and 101 of these patients died (stage 1a was absent, but there was 173 stage 1b events, 53 stage 2, and 15 stage 3 events in these 101 patients) while only 2 patients had AKI-events defined according to AKIN-4 but not according to KDIGO-4, none of these two have died. AKIN-4 defines significantly more 1a events compared to KDIGO-4. This happens because of the different time windows for stage 1b KDIGO-4 (7 days) compared to stage 1b AKIN-4 (48 hours). More precisely, in order to find patients who are in stage 1a, we define a lower bound which is ≥ 0.3 absolute SCr increase over a 48-hour window for both AKIN-4 and KDIGO-4; however, the upper is < 1.49 relative SCr increase over a 7-day window of observation for KDIGO-4 and < 1.49 relative

SCr increase over a 48-hour window of observation for AKIN-4. As a result, the exact same numbers for these definitions have not been calculated.

From Table 4.3, it can be seen that there are significantly more defined AKI events (overall) when KDIGO-4 (n = 2423) is compared to AKIN-4 (n = 1370), or when KDIGO (n = 2182) is compared to AKIN (n = 1169). The distribution over the different stages 1a, 1b, 2 and 3 also shows a significant difference (p $<$ 0.0001): 20.1% (= 486/2423), 47.0% (= 1138/2423), 24.1% (= 584/2423) and 8.9% (= 215/2423) for KDIGO-4 compared to 58.2% (= 797/1370), 30.8% (= 422/1370), 9.3% (= 127/1370) and 1.8% (= 24/1370) for AKIN-4. Using the modified KDIGO-4 the incidence of AKI was significantly higher for stage 1b compared to stage 1a (47.0% vs 20.1%, p $<$ 0.0001), while the opposite (30.8% vs 58.2%, p $<$ 0.0001) was observed when we used the modified AKIN-4 criteria. AKI stages 1a and 1b detect two different subgroups although there are differences between AKIN-4 and KDIGO-4, with AKIN-4 classifying more patients as 1a (58.2%) whereas KDIGO-4 more as 1b (47.0%).

Based on KDIGO-4, there were only 145 deaths out of 5713 patients (2.54%) that experienced 'no AKI' event, compared to 246 deaths out of 6172 patients (3.99%) that experienced 'no AKI' event based on AKIN-4 ($p < 0.0001$). These 145 deaths (without AKI-events according to KDIGO-4) were all part of the 246 deaths (without AKI-events according to AKIN-4), meaning that 101 deaths (= 246 – 145) experienced AKI-events as defined by KDIGO-4, but not by AKIN-4, on a total of 461 patients (21.9%). On the other hand, only 2 patients experienced AKI-events according to AKIN-4 but not according to KDIGO-4, and none of both died. Both definitions defined AKI events in 1068 patients, of whom 443 died (41.5%) (see Tables A.1 and Table A.2 in Appendix A).

AKI stages 1a and 1b accounted for 1624 cases defined by KDIGO-4 and 1219 cases defined by AKIN-4, and the mortality rate was 39% (= 633/1624) and 47% (= 572/1219) ($p < 0.0001$) respectively. The mortality rate in stages 1a and 1b was 38% and 39% ($p = 0.622$) respectively for KDIGO-4 while there was a substantial difference in mortality rate in stages 1a and 1b (43.9% vs 52.6%, $p = 0.004$) when based on AKIN-4. However, since in Table 4.3 patients that experienced both stages 1a and 1b were registered in both categories we analyzed stage 1 cases as patients who experienced only stage 1a, only stage 1b, and as patients who experienced both stages 1a and 1b. The results are shown in Appendix A Table A.3. Although there are about 5 times fewer patients in stages 2 & 3 (even 9 times less in stage 3) based on AKIN-4 (n = 151 in stages 2 & 3, n = 24 in stage 3) compared to KDIGO-4 (n = 799 in stages 2 & 3, n = 215 in stage 3), the mortality rate was about the same (55% vs 58%).

Table 4.3: Incidence of AKI and in-hospital mortality according to KDIGO-4, AKIN-4, KDIGO, and AKIN.

	Proportion of total patients meeting criteria		Incidence of in-hospital mortality	
KDIGO-4	n	%	n	%
No AKI	5713 out of 7242	78.56	145 out of 5713	2.54
Stage 1a	486 out of 7242	6.72	185 out of 486	38.06
Stage 1b	1138 out of 7242	15.71	448 out of 1138	39.35
Stage 2	584 out of 7242	8.06	327 out of 584	55.99
Stage 3	215 out of 7242	2.97	139 out of 215	64.65
Total AKI events	2423 out of 7242	33.45	1099 out of 2423	45.35
AKIN-4				
No AKI	6172 out of 7242	85.22	246 out of 6172	3.98
Stage 1a	797 out of 7242	11.00	350 out of 797	43.91
Stage 1b	422 out of 7242	5.83	222 out of 422	52.61
Stage 2	127 out of 7242	1.75	68 out of 127	53.54
Stage 3	24 out of 7242	0.33	15 out of 24	62.5
Total AKI events	1370 out of 7242	18.92	655 out of 1370	47.81
KDIGO				
No AKI	5708 out of 7242	78.82	145 out of 5708	2.54
Stage 1	1382 out of 7242	19.08	487 out of 1382	35.24
Stage 2	583 out of 7242	8.05	326 out of 583	55.92
Stage 3	217 out of 7242	2.99	140 out of 217	64.52
Total AKI events	2182 out of 7242	30.13	953 out of 2182	43.67
AKIN				
No AKI	6165 out of 7242	85.13	244 out of 6165	3.96
Stage 1	1018 out of 7242	14.06	424 out of 1018	41.65
Stage 2	127 out of 7242	1.75	68 out of 127	53.54
Stage 3	24 out of 7242	0.33	15 out of 24	62.5
Total AKI events	1169 out of 7242	16.14	507 out of 1169	43.37

4.3.3 Association between AKI events and mortality, based on multivariable models

Based on the number of AKI events/stages

Figure 4.1 shows the trend of mortality probability predicted by the logistic regression model with the number of AKI events for both KDIGO-4 and AKIN-4. There is a clear increasing trend of in-hospital mortality with the number of AKI events.

If we use a probability of 0.50 as the threshold to define "alive/dead", then according to AKIN-4, 3 to 4 events, or more, predict mortality, while according to KDIGO-4, 5 events or more are predictive for death. This is also reflected in the higher odds ratio for the number of AKI events in the logistic regression model when AKI events are defined by AKIN-4. Remember however that there are fewer AKI events (n = 1370) based on the AKIN-4 definition compared to the KDIGO-4 definition (n = 2423). In other words, the probability of dying for a fixed number of events (e.g., 5) based on both definitions, will be higher when the events are based on the AKIN-4 definition (prob = 75-90%), than based

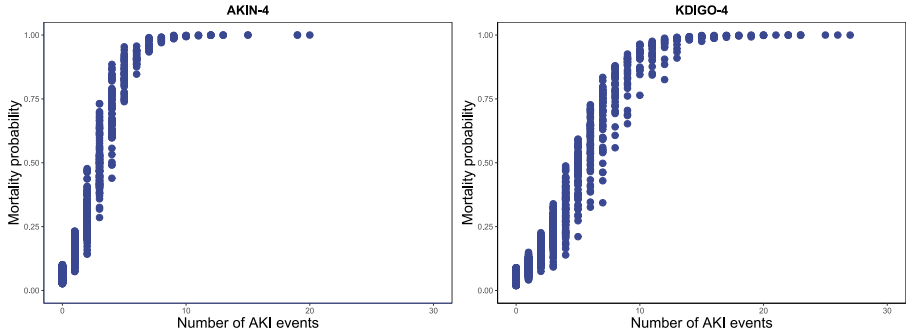


Figure 4.1: Mortality prediction using Logistic regression plotted versus the number of AKI-events.

on the KDIGO-4 definition (prob = 25-60%). We also investigated if there is a relation between mortality and the AKI profile (= the number of AKI events per stage, per patient). Results show that in-hospital mortality increased as the number of AKI events increased for both KDIGO-4 ($p < 0.001$) and AKIN-4 ($p < 0.001$) (Table 4.4). For every year older there is a 1.016 higher chance to die. Also, men have a 1.317 times higher chance to die than women. Moreover, for every stage 1a AKI event, the risk of death increases by a factor of 1.555 (as compared to no AKI event).

Table 4.4: Odds ratios (with 95% Confidence Interval) for the logistic regression models for in-hospital mortality.

OR [95%CI]	Number of AKI Events		Number of AKI Stages		Most Severe AKI Stage			
	KDIGO-4	AKIN-4	KDIGO-4	AKIN-4	KDIGO-4	AKIN-4	KDIGO	AKIN
Age	1.016*** [1.009-1.023]	1.014*** [1.008-1.020]	1.016*** [1.009-1.023]	1.015*** [1.008-1.021]	1.006* [0.999-1.013]	1.006* [1.000-1.013]	1.006* [0.999-1.012]	1.005* [0.999-1.011]
Sex (Male)	1.288* [1.056-1.569]	1.205* [0.998-1.455]	1.317** [1.078-1.608]	1.225* [1.013-1.480]	1.272* [1.042-1.552]	1.298** [1.075-1.567]	1.220* [1.002-1.486]	1.248* [1.036-1.504]
Stage 1a	1.690*** [1.626-1.759]	2.385*** [2.221-2.569]	1.555*** [1.331-1.824]	2.165*** [1.976-2.379]	4.689*** [2.978-7.156]	10.38*** [8.254-13.05]	9.725*** [7.702-12.30]	14.67*** [12.14-17.76]
Stage 1b			1.825*** [1.704-1.959]	3.093*** [2.614-3.686]	11.86*** [9.284-15.18]	22.75*** [17.78-29.18]		
Stage 2			1.677*** [1.522-1.855]	2.353*** [1.736-3.239]	39.79*** [30.61-51.96]	24.28*** [16.54-35.82]	39.72*** [30.55-51.87]	24.48*** [16.68-36.11]
Stage 3			1.405*** [1.237-1.603]	2.572** [1.488-4.779]	62.88*** [45.23-88.24]	37.45*** [16.06-94.04]	62.68*** [45.14-87.80]	37.59*** [16.14-94.33]

Significance codes : * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Based on the most severe stage of AKI

To find the association between AKI stages using KDIGO-4 and AKIN-4 and in-hospital mortality, we considered the most severe stage of AKI reached by a patient in a new logistic regression model. Results show that in-hospital mortality increased as the severity of AKI events increased for both KDIGO-4

($p < 0.001$) and AKIN-4 ($p < 0.001$) (Table 4.4). In addition, Figure A.4 in Appendix A shows the odds ratio of in-hospital mortality using logistic regression, stratified by the most severe stage of AKI events according to AKIN-4 and KDIGO-4. Finally, AKIN-4, overall, finds fewer AKI events based on the most severe stage, as seen in Table A.4 in Appendix A, AKIN-4 (1070) has a smaller number compared to KDIGO-4 (1529). Based on these results, we see that the mortality rate increases more gradually from stage 1a, 1b, 2, to 3 in the KDIGO-4 definition (12.6%, 24.6%, 50.7% to 64.7%) compared to the AKIN-4 definition (32.4%, 49.3%, 52.8% to 62.5%). Another fact is that the number of 'most severe AKI events' is very different between AKIN-4 and KDIGO-4, with much more events in stage 1a for AKIN-4 (n = 544) compared to KDIGO-4 (n = 222) while the inverse is true for stage 1b (n = 377 for AKIN-4 versus n = 667 for KDIGO-4). Finally, the events AKIN-4 detects for stage 2 and stage 3 are significantly lower compared to events detected by KDIGO-4. (125 vs 416 for stage 2 and 24 vs 215 for stage 3 respectively).

4.3.4 Comparing LR with RF for predicting mortality

Figure A.5 in Appendix A shows the ROC curve for using KDIGO-4 and AKIN-4, and KDIGO and AKIN for mortality prediction, respectively. By comparing ROC curves (or the AUCs from Table A.5 in Appendix A), we can conclude that using KDIGO-4 and AKIN-4 definitions have a slightly better (not significant) prediction compared to the original definitions (KDIGO and AKIN). Moreover, we found that logistic regression performs slightly better (not significant, $p < 0.005$) compared to random forest."

4.4 Discussion

4.4.1 AKI incidence by KDIGO-4 vs AKIN-4

The adoption of international criteria not only harmonized the definition of AKI, which is based on changes in SCr concentration and the degree of oliguria but also increased awareness and standardized the diagnosis of AKI. However, our data show that the combined diagnostic criteria in the KDIGO definition for stage 1 are not equivalent and that they detect as AKI stage 1 two distinct patient subgroups: one that is defined by an absolute increase of 0.3 mg/dL and one that is defined by the relative increase of 50% above the baseline.

In addition, our data show that KDIGO-4 and AKIN-4 definitions are very different, with KDIGO-4 being more sensitive compared to AKIN-4. These differences in the definitions clearly have consequences in terms of AKI incidence. For example, our analysis showed that a significant number of patients (461 patients) with AKI events defined by KDIGO-4 were classified as no-AKI

according to AKIN-4 and only 2 patients were diagnosed with the inverse verdict. These extra "no-AKI" cases defined by AKIN-4 exhibited significantly higher in-hospital mortality.

The most relevant difference between KDIGO and AKIN (in both standard and modified definitions) is related to the conditions necessary to classify patients and is the criterion that requires SCr to increase $> 50\%$ from baseline. Whereas AKIN requires this increase to happen within 48 hours the KDIGO requires this increase to happen within 7 days. It is obvious that in the strict time frame of 48 hours fewer patients will meet the required increase by AKIN compared to KDIGO where the timeframe is much wider. The longer we wait to observe the 50% increase in SCr, the more the sensitivity of the definition increases. This is also the case for stages 2 and 3 in both definitions where the incidence of AKI is significantly higher when the KDIGO definition is used compared to AKIN.

Moreover, based on our analysis the distribution of AKI events among the different stages (1a, 1b, 2, and 3) for both definitions are significantly different with KDIGO-4 defining significantly more 1b events compared to AKIN-4 and AKIN-4 to define significantly more 1a events compared to KDIGO-4. This happens because of the different time windows for stage 1b KDIGO-4 (7 days) compared to stage 1b AKIN-4 (48 hours). These findings support the conclusion that KDIGO-4 is more sensitive in detecting AKI events.

4.4.2 Impact of categorizing AKI stage 1 into stage 1a and stage 1b

AKI stage 1a represents patients whose reference SCr rises by 0.3 mg/dL, whereas AKI stage 1b represents patients whose reference SCr increases by 50%. Furthermore, our results show that these two criteria in KDIGO AKI stage 1 identify two different populations in terms of mortality. Table A.3 in Appendix A shows the number of patients experiencing only stage 1a, 1b, and both. While using KDIGO-4, the mortality rate for these subcategories of patients is significant (13%, 21%, and 43%), using AKIN-4 there are no significant differences between patients who experience only stage 1a and only stage 1b (32.35% and 32.33%). Furthermore, to find differences between these two subcategories, we also classified the patients by the most severe stage of AKI reached during the hospitalization (Table A.4 in Appendix A). The results show that the mortality rate among patients who experiences stage 1b as the most severe stage is two times higher than for the patients who experience stage 1a as the most severe stage. Consequently, the present study confirms that within the KDIGO AKI stage 1, there are two subpopulations with different severity of clinical outcomes (mortality). Additionally, patients with AKI stages 1a and 1b experienced clinically meaningful and statistically significant differences in

outcomes of in-hospital mortality (Table 4.4). This analysis demonstrates how different both definitions are, and also shows that separating stage 1 into 1a and 1b show a gradual increase in mortality rate.

4.4.3 Associations between AKI events and mortality

Based on the LR model, the odds for in-hospital mortality were progressively higher for patients with AKI compared to patients without AKI, and it was higher with higher stages. This was evident with both definitions: AKIN-4 and KDIGO-4. Moreover, the odds for in-hospital mortality were positively associated with the number of AKI events of the patient. Results show that when predicting adverse outcomes (in-hospital mortality in our case), classification seems better with KDIGO and KDIGO-4 systems. Additionally, our results show that due to lack of sensitivity, the AKIN-4 definition classifies more cases as "no-AKI" compared to KDIGO-4. These "no-AKI" cases exhibited significantly higher mortality during the observation period (22.0% incidence of in-hospital mortality). This explains the increased overall incidence of mortality observed among "no-AKI" cases as defined by AKIN-4 compared to KDIGO-4. In addition, KDIGO (and KDIGO-4) classifies more patients as stage 2 and stage 3 than AKIN (and AKIN-4). These findings support the conclusion that the classification of a patient at a higher stage of AKI with all definitions (in both standard and modified definitions) has a progressively larger negative impact on mortality. However, KDIGO-4 (KDIGO) has a higher odds ratio at a higher stage of AKI compared to AKIN-4 (AKIN). Moreover, this study demonstrates that KDIGO-4 and AKIN-4 definitions act differently in detecting AKI events, and also shows that separating stage 1 into 1a and 1b has clinically meaningful and statistically significant differences in the outcome of in-hospital mortality.

4.5 Conclusion

This study demonstrates that KDIGO-4 and AKIN-4 definitions act differently in detecting AKI events, and also shows that separating stage 1 into 1a and 1b has clinically meaningful and statistically significant differences for outcomes of in-hospital mortality. Repeated AKI episodes are also associated with mortality. In addition, results confirm a higher stage of AKI with all definitions (in both standard and modified definitions) has a progressively larger negative impact on mortality.

Acknowledgments

The authors would like to thank the patients and staff of KAT General Hospital without whom this research would not have been possible.

Conflict of interest statement

The authors of this paper have no conflicts of interest to disclose.

Author contributions

KM and MA designed the study. FNH and HP performed the statistical analysis. FNH and KM wrote the first draft of the manuscript with input from HP, PD, and EC. The manuscript was reviewed and approved by all authors before submission.

Funding

The authors of this manuscript received no financial support for the research, authorship, or publication of this article. This research was not supported by any funding.

Data availability statement

The data supporting the results of this study cannot be made publicly available due to the lack of approval from our ethics committee in this regard.

Chapter 5

Validated risk prediction models for outcomes of acute kidney injury

The following chapter has been submitted to BMC Nephrology:

Nateghi Haredasht, F., Vanhoutte, L., Vens, C., Pottel, H., Viaene, L., & De Corte, W. Validated risk prediction models for outcomes of acute kidney injury: a systematic review.

Abstract

Background: Acute Kidney Injury (AKI) is frequently seen in hospitalized and critically ill patients. Studies have shown that AKI is a risk factor for the development of acute kidney disease (AKD), chronic kidney disease (CKD), and mortality.
Methods: A systematic review is performed on validated risk prediction models for developing poor renal outcomes after AKI scenarios. Medline, EMBASE, Cochrane, and Web of Science were searched for articles that developed or validated a prediction model. Moreover, studies that report prediction models for recovery after AKI also have been included. This review was registered with PROSPERO (CRD42022303197).
Result: We screened 25812 potentially relevant abstracts. Among the 149 remaining articles in the first selection, eight met the inclusion criteria. All of the included models developed more than one prediction model with different variables. The models included between 3 and 28 independent variables and c-statistics ranged from 0.55 to 1.
Conclusion: Risk prediction models for developing renal insufficiency after experiencing AKI are based on simple statistical/machine learning models. We believe that advanced machine learning models using big data information are required to increase the predictive performance for developing renal insufficiencies.

5.1 Introduction

Acute kidney injury (AKI) among hospitalized patients is characterized by a sudden decline in renal function and is associated with poor long-term and short-term outcomes [9]. The overall incidence of AKI in hospital patients ranges between 7% to 22%, and it ranges from 20% to 50% in the Intensive Care Unit (ICU) patients [201, 20]. It has been shown that when sepsis is present at ICU admission, the prevalence of AKI is greater than 40% [4].

The definition of AKI has changed over the years. In 2012, the Kidney Disease: Improving Global Outcomes (KDIGO) unified the previous definitions (RIFLE and AKIN) [92]. By KDIGO definition, AKI is diagnosed by an absolute increase in SCr, at least $0.3mg/dL$ ($26.5\mu mol/L$) within 48 hours or by a 50% increase in SCr from baseline within 7 days, or a urine volume of less than $0.5mL/kg/h$ for at least 6 hours.

Although KDIGO is now the most accepted and used AKI criteria, recently Sparrow et al. [177] evaluated the impact of further sub-categorizing the KDIGO-defined AKI stage 1 into two stages based on SCr criteria: stage 1a (an absolute increase of SCr of 0.3 mg/dL within 48 h) and stage 1b (a 50% relative increase in SCr within 7 days) and therefore creating a 4-stage KDIGO classification which they named KDIGO-4. In a separate study, Nateghi et al. [133] showed that within the KDIGO AKI stage 1, there are indeed two sub-populations with different clinical outcomes.

AKI contributes to adverse short-term and long-term outcomes. Different studies have linked AKI to the development of acute kidney disease (AKD) (see Table 1.2), chronic kidney disease (CKD) see Table 1.2), end-stage kidney disease, longer hospitalization time, cardiovascular disease (CVD), and other complications, suggesting that even a short episode of acute kidney injury might lead to long term morbidity [178] and mortality [132],[172]. Among the 19,249 hospitalizations included in a study in which the incidence of AKI was 22.7%, Wang et al. [201] reported the mortality rate was 10.8%, compared to 1.5% for cases without AKI. Moreover, it has been reported that critically ill patients with dialysis-requiring AKI experience mortality rates above 50% [137]. The mortality rate of this sudden kidney failure in ICU is approximately 30%–50% depending on the medical record of the patient and the stage of AKI [8, 62].

Traditionally, most studies of severe AKI have concentrated on short-term outcomes often evaluated at hospital discharge. However, AKI may exhibit important independent effects on the outcome that may extend well beyond discharge from the hospital [52]. Figure 5.1 shows the potential long-term outcomes in AKI. As a result of an episode of AKI, patients may recover, be discharged without recovery of renal function, or die. Patients who seem to recover may also later develop CKD or CVD.

In recent years, it has become clear that AKI is not a completely reversible syndrome. It is possible that the injury that occurs may result in permanent kidney damage (e.g., CKD) and even damage to other organs.

This caused a shift from AKI being a life-threatening and acute situation to a situation with a larger population in need of chronic follow-up to prevent further deterioration of their kidney function [108].

AKI and CKD have been associated, however, this can be explained by confounding factors and bias, thereby questioning the causal nature of the findings [80]. Nevertheless, in light of the association and the increasing number of patients with AKI (so-called AKI survivors), and CKD, the prediction of CKD after an AKI episode has become increasingly crucial in order to allocate the necessary amount of follow-up to the right patients.

Currently, follow-up of AKI survivors is often lacking and not regulated: follow-up of kidney function by a nephrologist in patients surviving an episode of AKI treated with renal replacement therapy (RRT) is stated in nearly one-third of the patients [24]. Close follow-up and interventions aimed at preserving kidney function may positively impact long-term outcomes. However, this is costly and time-consuming. As a result, instead of monitoring all the patients experiencing AKI, it would be useful to identify those subgroups of patients who are at higher risk of developing CKD and only follow up with those patients. In order to do so, we need to collect data to be able to develop a prediction model to output a risk score for developing CKD for patients who experienced AKI.

Lately, with the help of technology, e.g., electronic health records (EHR), collecting clinical and biochemical data is much more straightforward than before [15]. As a result, the resulting data could be analyzed and prediction models could be constructed. Recently, there have been several studies using machine learning technology for outcome prediction using EHR data [144, 91]. One of the main tasks considered in machine learning is the development of a model by learning from a set of observed data in order to predict outcomes or events for future data [19]. Although the traditional statistical approaches appear to be more appropriate when a large number of cases exceed the number of variables under study and significant a priori knowledge of the subject area is available, machine learning algorithms can handle a large amount of data with high-dimensional variables. In addition, interpretable machine learning models make it possible for healthcare experts to make individualized decisions that will eventually lead to a higher standard of care.

Objective In this systematic review with meta-synthesis, we investigate the use of validated predictive models (machine learning or statistical models) for predicting the development of renal insufficiency in the short-term and long-term after AKI scenarios in the hospital/ICU. The purpose of a meta-synthesis is

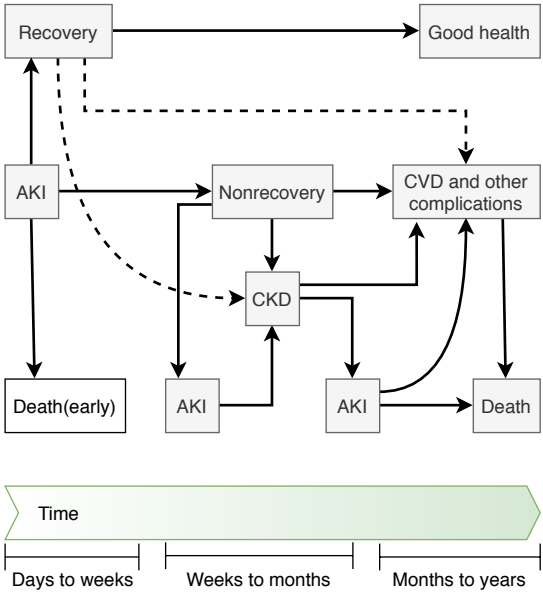


Figure 5.1: Possible outcomes following AKI. As a result of an episode of AKI, patients may recover, be discharged without recovery of renal function, or die. Patients who seem to recover may also later develop CKD or CVD (dashed lines)- modified from reference [93].

to synthesize qualitative data in order to provide a new interpretation of the research field. In contrast to meta-analyses, which use quantitative data to test a hypothesis, it helps to build new theories. Since it is essential to assess the degree to which a model generalizes, we focused specifically on models that have been validated either externally (e.g., separate cohort) or internally (e.g., cross-validation). Validating a prediction model plays a particularly important role in the healthcare domain since the ultimate purpose of developing a model is to use it in clinical settings, and providing a validated mode enhances its reliability.

5.2 Materials and methods

Published guidance (CHARMS, TRIPOD, and Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA)) helped frame the review question, data extraction, reporting, and appraisal. The protocol of our systematic review has been previously registered at the PROSPERO International Prospective Register of Systematic Reviews website (under the reference CRD42022303197).

Table 5.1: Search strategy: keywords and MeSH terms for systematic literature review in Pubmed.

Concept	Keywords *	MeSH terms
1. Acute Kidney Injury	"acute kidney injur*", "acute renal injur*, "acute renal insufficienc*, "acute kidney insufficienc*, "acute kidney failure*", "acute renal failure*", "renal insufficienc*", "kidney insufficienc*", "kidney dialys*", "renal dialys*", "hemodialys*", "hemodiafiltration"	"acute kidney injury", "renal insufficiency"
2. Outcome of AKI	"chronic renal insufficienc*", "chronic kidney insufficienc*", "chronic kidney disease*", "chronic renal disease*", "end-stage kidney disease*", "end-stage renal disease*", "end-stage kidney failure*", "chronic kidney failure", "chronic renal failure", "ESRD", "follow-up stud*", "cohort stud*", "cohort analys*, "follow-up", "long-term outcome*"	"renal insufficiency, chronic", "kidney failure, chronic", "follow-up studies", "cohort studies",
3. AI/machine learning	"artificial intelligence","machine intelligence", "computational intelligence", "statistical model*", "probabilistic model*", "decision support technique*", "decision support model*", "decision support system*", "decision analys*", "decision model", "predict model*", "prediction model*", "predict rule*", "predict score", "prediction score*", "prognostic model*", "decision rule", "risk model*", "risk algorithm*", "validation", "risk index", "risk predict*", "clinical model*" "survival analysis", "proportional hazard model*", "Kaplan-Meier survival curve", "cox model*, "time-to-event analysis", "machine learning", "transfer learning", "deep learning", "supervised machine learning", "learning from labeled data", "logistic model*"	"artificial intelligence", "models, statistical", "decision support techniques", "survival analysis", "risk"

* Throughout the table, * is a truncation symbol.
Searches combined with AND: 1 AND 2 AND 3. The same search query has been adapted to be used in Web of Science, Cochrane, and Embase.

5.2.1 Search strategy

We searched Medline, EMBASE, Cochrane, and Web of Science for review articles and regular research articles, from January 1st, 2011 to January 12th, 2022. Apart from restricting English language articles, no further restrictions were applied. Three search themes were used in the query: "acute kidney injury", "outcome of AKI", and "artificial intelligence". We also adapted these keywords to Medical Subject Heading (MeSH) terms according to the CHARMS guideline. To ensure consistency in the searches for all databases, first, we set up the search in Pubmed, then the query was translated to EMBASE, Cochrane, and Web of Science. Table 5.1 shows our search strategy with every keyword and detail.

5.2.2 Selection criteria

The purpose of this section is to discuss our criteria for including and excluding articles, and the steps taken by the reviewers to determine which articles were included or excluded.

Inclusion Two independent reviewers (FNH and LV) screened all titles and abstracts identified by querying the databases using the search strategy detailed above. Articles identified as potentially relevant by either reviewer were subsequently read in full. Full-text articles were included if they (i) developed a machine learning-based or statistical prediction model for predicting renal insufficiency after an episode of AKI, and (ii) assessed the impact of the predictive model for renal insufficiency after an episode of AKI that was implemented in a clinical setting.

Exclusion In this phase of the selection, articles were excluded based on the following criteria: (i) not a prediction model study, (ii) renal insufficiency is not the outcome, and (iii) no validation of the model (neither internal nor external).

5.2.3 Data extraction

The same two reviewers extracted data from the articles using a meticulously composed data extraction form that was designed in advance. The acquired data consists of: (i) the study setting, (ii) derivation and validation cohort descriptions, (iii) modeling approach, (iv) validation method, (v) model performance statistics, and (vi) final prediction tool design. We allowed details of external validation to be included in the extracted data when they were part of a preceding or sequential publication.

5.2.4 Model performance

We gathered information concerning model discrimination and calibration using multiple units or by a combined measure, in order to evaluate the models' performance. Calibration refers to the agreement between observed outcomes and predictions meaning that in this context if a model predicts a 40% risk of developing renal insufficiency for an AKI patient, the observed frequency of renal insufficiency should be approximately 40 out of 100 AKI patients with such a prediction [68]. The assessment of calibration consists of evaluating whether predicted probabilities and observed probabilities agree, including goodness-of-fit tests [for example, Hosmer–Lemeshow (HL) tests], table or graphical comparisons of predicted versus observed values within groups of predicted risks, or calibration plots. Poor calibration is indicated by an HL statistic with a small, significant p-value. Accordingly, discrimination is defined as the ability to distinguish between patients who are likely to develop renal insufficiencies

such as acute kidney disease (AKD), which is a condition that falls between AKI and chronic kidney disease (CKD) and patients who are likely to develop CKD following an episode of AKI. Discriminating power was assessed using the area under the receiver operating characteristic (AUROC)/c-statistics [60]. Any information about the matching of model-predicted probabilities and observed probabilities was also included in the assessment of model performance, for example, the goodness-of-fit test, Hosmer-Lemeshow test [71], or table/graphical visualization of prediction versus observation values/performance.

5.2.5 Study quality assessment

An assessment of quality criteria was conducted based on the Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD statement) [30]. There is no standardized mechanism to assess the quality of impact analysis studies for risk prediction models. Therefore, quality criteria have been adapted from published articles that address the validity of prediction models in clinical implementation and impact analysis phases [65, 207].

5.3 Results

5.3.1 Characteristics of the included studies

We identified 33746 potentially relevant abstracts from the searches over all of the databases. We also found one study from other sources and references. After the duplicate removal, as well as 25812 title/abstract screening, 149 studies were assessed for full-text review. After full article screening, eight articles were identified for information extraction. As a result, we reviewed eight studies that reported prediction models. Figure 5.2 shows the flow of articles based on our search strategy. A summary of the predictive variables included in models is found in Table 5.2.

Chawla et al. [24] conducted a prospective single-center cohort study in which they developed three prediction models to identify patients who survive AKI and are at higher risk for progression to stage 4 CKD. First, a model using all variables was developed, then a stepwise forward selection procedure with a threshold of $P < 0.1$ was used for feature selection. Then a second model was developed using the most heavily weighted factors from the first model. Following that, a third model was developed, called the 'bedside' model, which is based on sentinel clinical events. The prediction models were built using multivariate logistic regression methodology. They validated the three models externally using a separate cohort. In the model validation on the test set (separate validation cohort), models 1, 2, and 3 were all statistically significant

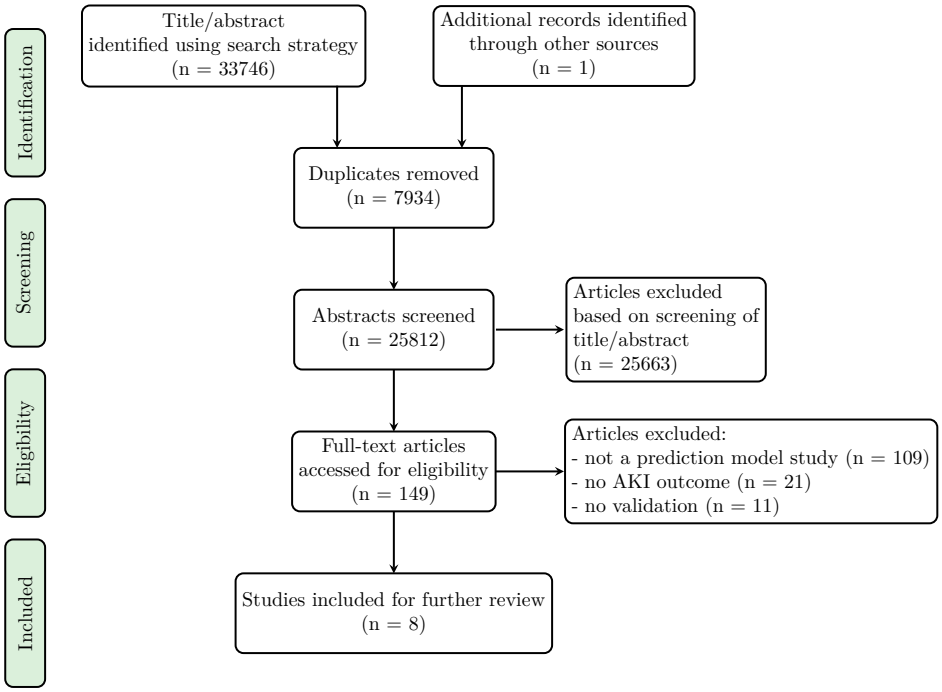


Figure 5.2: Flow of articles using our search strategy.

in predicting progression to stage 4 CKD with c-statistics of 0.82, 0.81, and 0.77, respectively ($P < 0.05$ was the level of significance).

Itenov et al. [85] performed a multi-center prospective study on a cohort of adult critically ill patients admitted to the ICU for at least 24 hours and with AKI defined by KDIGO. The main outcome of this study was a recovery of kidney function within 28 days in which recovery is defined as living for five consecutive days with no renal replacement therapy and with creatinine levels below 1.5 times the baseline value (measured before ICU admission). The prediction models, one of which is for the hazard of recovery and the other for death without recovery, were developed by combining cause-specific Cox regression [34] models. Two separate prediction models were developed using different variables. The "basic" model has the most likely variables including elevation in creatinine, urinary output, sex, and age. The second model, the "full" model, incorporates all the potential predictors (see Table 5.2). The models were validated on a separate validation cohort showing that 59.1% of the patients recovered, meaning that almost 40.9% of the patients developed any kind of renal insufficiency (e.g., different stages of CKD). In addition, 9.0%

Table 5.2: Predictive variables included in the models.

Variable	Chawla, L.S. et al (2011)	Itenov, T.S. et al (2018)	James, M.T. et al (2017)	Lee, B.J. et al (2019)
Demographics				
Age	✔	✔	✔	✔
Gender/Sex	Male/Female	Female	Male	✘
Race	African American/Hispanic/ Caucasian/Other	✘	✘	✘
Laboratory data				
Baseline serum creatinine, mg/dL	✘	✘	✔	✘
Serum creatinine, mg/dL	✔	✘	✘	✘
Discharge serum creatinine, mg/dL	✘	✘	✔	✘
Delta creatinine, mg/dL	✘	✔	✘	✘
Urinary output, $mL/kg/h$	✘	✔	✘	✘
Delta urinary output, $mL/kg/h$	✘	✘	✘	✘
Baseline eGFR, $mL/min/1.73m^2$	✔	✘	✘	✘
Interleukin-8	✘	✘	✘	✘
Interleukin-16	✘	✘	✘	✘
AKI stage	✘	✘	1/2/3	✘
Albuminuria	✘	✘	Normal/Mild/Heavy/Unmeasured	✘
Baseline serum albumin (Alb)	✔	✘	✘	✘
Serum albumin (Alb)	✔	✘	✘	✘
Baseline serum hemoglobin (Hgb)	✔	✘	✘	✔
Serum hemoglobin (Hgb)	✔	✘	✘	✘
Total bilirubin	✘	✘	✘	✘
Maximum urea before first AKI-3	✘	✘	✘	✘
Maximum white blood cell count before first AKI-3	✘	✘	✘	✘
Preadmission platelet count, $\times 10^3/\mu l$	✘	✘	✘	✘
Comorbidities				
Apache II score	✘	✘	✘	✘
Oliguria	✘	✘	✘	✘
Mechanical ventilation	✘	✘	✘	✘
Diabetes mellitus (DM)	Yes/No	✘	✘	✘
Dialysis	Never/During hospitalization/ Post hospitalization	✘	✘	✘
Chronic liver disease	✘	✘	✘	Yes/No
Renal replacement therapy (RRT)	✔	✘	✘	✘
Arterial pH (Z-score)	✘	✘	✘	✘
Platelets	✘	✘	✘	✘
Mean arterial pressure	✘	✘	✘	✘
Acute tubular necrosis	Yes/No	✘	✘	✘
Time at risk (years)[1]	Yes/No	✘	✘	✘
Hospital complexity	1A/1B/1C/2/3	✘	✘	✘
Residency slots	✔	✘	✘	✘
Teaching hospital[2]	Yes/No	✘	✘	✘
Sepsis	✘	✘	✘	✘
Mechanical ventilation	✘	✘	✘	✘
Chronic obstructive pulmonary disease	✘	✘	✘	✘
APS III score	✘	✘	✘	✘
Diabetes	✘	✘	✘	✘
Congestive heart failure	✘	✘	✘	✘
Moderate or severe liver disease	✘	✘	✘	✘
SAPS II score	✘	✘	✘	✘
SOFA score	✘	✘	✘	✘
RRT on the first AKI-3 day in ICU	✘	✘	✘	✘
Hypertension	✘	✘	✘	✘
Surgery/trauma	✘	✘	✘	✘
Diuretic	✘	✘	✘	✘
Renal toxic drugs	✘	✘	✘	✘
Charlson Comorbidity Index	✘	✘	✘	✘
Emergency department	✘	✘	✘	✘

[1] Years from diagnosis date to either the end of the data collection period or date of death, whichever came first.

[2] Teaching hospital was coded yes when the number of Medical Residents was ≥ 5.

had a predicted chance of recovery of less than 25%, and their observed rate of recovery was 21.5%. The AUROC curve (or equivalently, the c-statistic) for predicting a recovery in the validation cohort was 73.1%.

James et al. [86] performed a multi-center prospective study in which they derived and internally as well as externally validated five different predictive models for the progression of AKI to advanced chronic kidney disease. Candidate predictor variables were selected based on previous studies. Then, stepwise backward variable selection with a significance level of $P < 0.05$ was used for the feature selection procedure. Five models with different variables were developed using multivariate logistic regression and internally validated using a random

Table 4 continued:: Predictive variables included in the models.

Variable	Chen, Z. et al (2021)	He, J. et al (2021)	Huang, C.Y. et al (2022)	Pike, F. et al (2015)
Demographics				
Age	✘	✔	✔	✔
Gender/Sex	✘	✔	✘	✘
BMI, kg/m^2	✘	✔	✘	✘
Laboratory data				
Baseline serum creatinine, mg/dL	✘	✔	✘	✘
Serum creatinine, mg/dL	✘	✔	✘	✘
Delta creatinine, mg/dL	✔	✔	✘	✘
Urinary output, $mL/kg/h$	✘	✔	✘	✘
Delta urinary output, $mL/kg/h$	✘	✔	✘	✘
Baseline eGFR, $mL/min/1.73m^2$	✘	✘	✘	✘
Interleukin-8	✔	✘	✘	✔
Interleukin-16	✔	✘	✘	✘
AKI stage	✘	1/2/3	✘	✘
Albuminuria	✘	✘	✘	✘
Baseline serum albumin (Alb)	✘	✘	✘	✘
Serum albumin (Alb)	✘	✘	✘	✘
Baseline serum hemoglobin (Hgb)	✘	✘	✘	✘
Serum hemoglobin (Hgb)	✘	✘	✘	✘
Total bilirubin	✘	✘	✘	✔
Maximum urea before first AKI-3	✘	✘	✔	✘
Maximum white blood cell count before first AKI-3	✘	✘	✔	✘
Preadmission platelet count, $\times 10^3/\mu l$	✘	✘	✔	✘
Comorbidities				
Apache II score	✘	✘	✘	✔
Oliguria	✘	✘	✘	✔
Mechanical ventilation	✘	✘	✘	✔
Diabetes mellitus (DM)	✘	✘	✘	✘
Dialysis	✘	✘	✘	✘
Chronic liver disease	✘	✘	✘	✘
Renal replacement therapy (RRT)	✘	✘	✘	✘
Arterial pH (Z-score)	✘	✘	✘	✔
Platelets	✘	✘	✘	✔
Mean arterial pressure	✘	✘	✘	✔
Acute tubular necrosis	✘	✘	✘	✘
Time at risk (years)	✘	✘	✘	✘
Hospital complexity	✘	✘	✘	✘
Residency slots	✘	✘	✘	✘
Teaching hospital	✘	✘	✘	✘
Sepsis	✘	✘	Yes/No	✘
Mechanical ventilation	✘	✔	✘	✘
Chronic obstructive pulmonary disease	✘	✔	✘	✘
APS III score	✘	✔	✘	✘
Diabetes	✘	✔	✘	✘
Congestive heart failure	✘	Yes/No	✘	✘
Moderate or severe liver disease	✘	Yes/No	✘	✘
SAPS II score	✘	✔	✘	✘
SOFA score	✘	✔	✘	✘
RRT on the first AKI-3 day in ICU	✘	✘	✔	✘
Hypertension	✘	Yes/No	✘	✘
Surgery/trauma	✘	Yes/No	✔	✘
Diuretic	✘	Yes/No	✘	✘
Renal toxic drugs	✘	✔	✘	✘
Charlson Comorbidity Index	✘	✔	✘	✘
Emergency department	✘	✔	✘	✘

2-fold split technique as well as externally on an independent cohort. Out of five different models, the first model (6-variable model) had the highest c-statistic of 0.87 (95% CI) and 0.81 (95% CI) in the internal and external validation cohort, respectively.

Lee et al. [104] published a multi-center retrospective cohort study on a cohort of dialysis-requiring adult acute kidney injury (AKI-D) patients who had predicted inpatient mortality of $< 20\%$. The study aimed to develop and validate a prediction model for the probability of recovery in these patients. Different candidate predictors including demographic characteristics, comorbidities, laboratory values, and medication were used to develop two

models using logistic regression and classification and regression tree (CART). Both models were internally validated using 10-fold cross-validation. Predicted recovery probabilities ranged from 9% to 22% in the lowest decile to 58% to 66% in the highest decile for logistic regression, and from 25.6% to 52.7% for the CART approach. The c-statistic was 0.64 and 0.61 for logistic regression and CART techniques, respectively.

Pike et al. [147], reported a multi-center prospective cohort study aiming to develop a biomarker-enhanced risk prediction model for critically ill patients receiving RRT with AKI. They investigate whether plasma inflammatory and apoptosis biomarkers increase risk prediction of renal recovery and mortality compared with clinical models in which the primary outcomes of interest were renal recovery and mortality at day 60. Four different models were developed using multivariate logistic regression in which each model uses a different set of variables (see table 5.3). The results of these four models were compared to an already existing model named Acute renal failure Trial Network (ATN) study [143] which uses 21 variables. As validation, they randomly split the dataset into two sets named derivation and validation sets. The c-statistic for all biomarkers for recovery and mortality were 0.66 and 0.71, respectively. The results show that a simple four-variable clinical model including age, mean arterial pressure, mechanical ventilation, and bilirubin, together with IL-8, increases prediction quality for renal recovery and mortality at day 60 and could potentially be beneficial at the bedside for clinicians.

A separate study conducted by Chen et al. [28] analyzed 32 immunoinflammatory cytokines in the blood of patients with cardiac surgery-associated acute kidney injury (CSA-AKI) and then employed machine learning methods to develop a simple and effective blood marker-based model for predicting poor in-hospital outcomes. CSA-AKI, defined as abrupt renal dysfunction that occurs in patients following cardiac surgery, is a prevalent complication affecting approximately 5 percent to 42 percent of patients undergoing cardiac surgery [69]. Using both Least Absolute Shrinkage and Selection Operator (LASSO) and random forest predictor selection methods, they showed a logistic regression-based predictive model incorporating IL-8, IL-16, and a change in SCr assists in accurately predicting poor in-hospital outcomes. The generalizability of the proposed models was tested using both internal validation (bootstrap) and external validation (separate cohort). Their prediction model was effective at predicting composite outcomes, reporting AUROC of 0.947 and 0.971 for internal and external validation, respectively.

In a separate study that studied the outcome in critically ill patients with sepsis-associated AKI, He et al. (2021) [66] developed and validated machine learning models to predict the occurrence of AKD [43]. AKD was defined as the presentation of at least KDIGO Stage 1 criteria for > 7 days after an

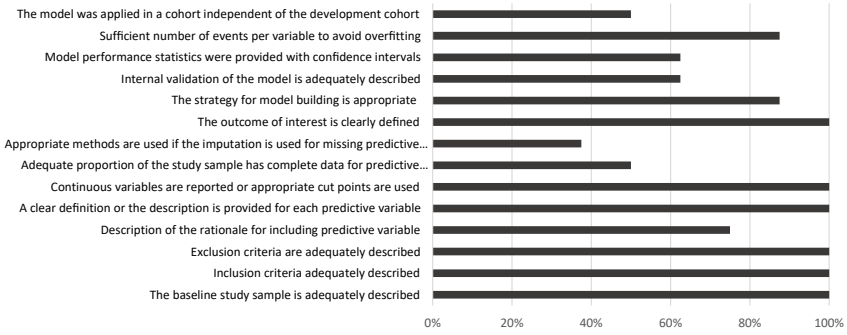


Figure 5.3: Percentage of studies meeting quality criteria.

AKI-initiating event [25]. To determine the most useful predictive variables, LASSO has been used and 28 variables (listed in Table 5.2) have been selected for inclusion in the predictive models. The results of three different models, including recurrent neural network-long short-term memory (RNN-LSTM), decision tree, and logistic regression, were compared on two separate training and validation (MIMIC III) datasets. In the validation dataset, the RNN-LSTM algorithm showed the highest performance with an AUROC of 1.000, followed by the decision trees with an AUROC of 0.872. Logistic regression had the least predictive accuracy, with an AUROC of 0.717.

Recently, Huang et al. [82] developed and validated prediction models for AKI recovery in critically ill patients at hospital discharge with ICU-acquired AKI stage 3 (AKI-3). After internal (10-fold cross-validation) and external validation the prediction LASSO model for complete or partial recovery based on age, need for RRT, platelet count, urea, and white blood cell count had the highest AUROC of 0.61.

A comparative summary of all clinical prediction models is shown in Table 5.3 and a summary of their methodological quality is provided in Figure 5.3.

5.3.2 Quality assessment summary

Table 5.4 shows the quality assessment of model development of the included studies. Overall, all studies met most quality measures. In addition, all studies except the ones by Chawla et al. [24] and He et al. described the rationale for including predictive variables. However, only three studies by Chawla et al. [24], Huang et al. [82], and Pike et al. [147] discussed handling missing data. The number of events per variable was < 10 for the study conducted by He et al. [66], and four of the eight models were validated externally.

5.4 Discussion

In this systematic review, we aimed to find prediction models for the development of renal insufficiency (or recovery) in patients who experienced AKI. We identified eight studies in which multiple prediction models were built and validated in heterogeneous cohorts of patients. AKI was defined using the KDIGO criteria in four studies [85, 86, 66, 82], and one study used the RIFLE criteria [24], the other three studies did not mention the used AKI criteria [104, 147, 28]. Our systematic review found some limitations in the derivation and validation of all published studies.

For a model to be generalizable beyond a sample population, validation is an essential step. Although all the models underwent some internal validation and reported model calibration (except Chawla et al. [24]), not all of them were externally validated. In addition, internal validation in one of the studies was performed in a random split of the dataset [147], which is not a perfect method for data splitting in that it generates quite similar development and validation set.

While some studies did not mention how missing values were handled, of those that did, the majority relied on relatively simple methods, such as complete case analysis and single imputation using mean for continuous data and the mode for categorical data. Only one study used a regression-based algorithm [147]. Multiple imputation methods have proven to be more effective than single imputation methods at restoring the natural variability of missing values and retaining more useful information than complete case analysis methods [181].

Moreover, three of the studies selected risk factors using LASSO for variable selection [66, 28, 147]. However, four of the eight models used statistical approaches of forward selection or backward elimination [24, 86, 104, 147], and one used correlation-based techniques [82]. Studies conducted using stepwise regression techniques have demonstrated wide variation in models selected from a list of candidate predictors. By bootstrapping for predictor selection, model developers can take into account this variability since the final candidate predictors are those selected by a predetermined majority of bootstrap samples. Only one model was developed using a full model approach.

In addition, three of the studies only focused on one particular center [24, 28, 66]. Using a single-center cohort may not be representative of other populations. In addition, all models were derived and validated in cohorts from the same region

meaning that generalizability to patients from other regions was not examined. Moreover, all studies excluded patients with preexisting CKD, therefore these prediction models may not be accurate in that population.

In the included studies, conventional statistical models or simple machine learning techniques such as CART, RNN, and logistic regression were the methods employed in this area. Rajula et al. [153] showed the traditional statistical method seems more useful than machine learning models when the number of cases is greater than the number of variables when applied to the medical field. However, in scenarios where the number of variables is large, traditional statistical models might run into problems. EHRs are capable of storing a large number and variety of variables enabling high-quality and trustworthy prediction models [94], and machine learning offers the techniques to handle large amounts of high-dimensional data where the number of variables is huge that is common in healthcare settings. Besides, these machine learning models are capable of capturing complex interactions between the variables in the datasets, resulting in more precise and reliable models. However, statistical models that leverage the diversity and abundance of EHR-derived data are still limited. Furthermore, many machine learning models like random forest [17] are able to handle missing values (one of the main challenges when developing EHR-based models) naturally, without the need to include a data imputation step. Also, the interpretability of model predictions is an important consideration when implementing and utilizing them by clinical providers and other healthcare decision-makers, and some machine learning models such as decision trees and random forests can be more easily interpreted. Despite many advantages, most machine learning models (e.g. deep learning) are computationally expensive and need more time for training.

Despite the fact that hyperparameter selection can greatly influence the performance of a model, hyperparameter selection is often neglected in these studies [117]. It is our understanding that there are no guidelines regarding how to report the hyperparameter tuning results/procedure for machine learning as clinical prediction models.

Another important issue is the limited amount of follow-up data. Based on the results of included papers, the need for early detection and prevention of AKI is important. However, currently, after discharge from the hospital, the follow-up of AKI survivors is considerably challenging mainly due to two reasons. First, the process is time-consuming and costly, and second, drop-out is frequently observed [123]. As a result, when developing machine learning-based CKD risk prediction models for such patients, we are typically confronted with a small labeled training set. For future research, we propose organizing longer follow-up studies of AKI patients, utilizing advanced machine learning methods

to take into account as many variables as possible, and employing techniques of semi-supervised learning to deal with probable dropouts [135].

It is important to note that this systematic review has both strengths and limitations. This is the first systematic review to examine both the reporting quality and the development of machine learning models that predict outcomes of AKI. Although we used standard search filters for AKI, outcomes of AKI, and machine learning, we may not have found all relevant studies in the databases that we have looked into or studies that are not included in these databases and not published in English. Moreover, it was not possible to perform a meta-analysis of the studies because access to individual participant data was not available. Finally, an individual model cannot be recommended or implemented due to the limited number of externally validated models and the absence of an impact analysis.

5.5 Conclusion

In recent years, few validated clinical models have been developed that can predict the outcomes of acute kidney injury in critically ill or hospitalized patients. Machine learning models were used in a limited number of applications and most of the models were based on traditional statistical models. Although some of these models were externally validated, none of these models are available in a manner that could be utilized or evaluated in clinical settings. It is not possible to say with confidence whether machine learning is superior to traditional statistical methods since a significant amount of development and improvement can still be made in machine learning prediction algorithms in this area, including utilizing a broader range of data sources, improving model selection and design, reporting the development process, and providing the source code of the final model for public access to be used in clinical settings. The existence and use of such models, in addition to highlighting increased renal insufficiency, morbidity, and mortality following AKI, have significant implications for the future care needs of survivors.

Acknowledgements

This work was supported by KU Leuven Internal Funds (grant 3M180314). The authors also acknowledge the Flemish Government AI Research Program. The authors wish to thank Thomas Vandendriessche, Kristel Paque, and Krizia Tuand, the biomedical reference librarians of the KU Leuven Libraries – 2Bergen – learning Centre Désiré Collen (Leuven, Belgium), for their help in conducting the systematic literature search.

Funding

This work was supported by KU Leuven Internal Funds (grant 3M180314).

Conflict of Interest Statement

None declared. The results presented in this paper have not been published previously in whole, part, or in abstract form.

Table 5.3: AKI-outcome prediction models.

	Chawla, L.S. et al (2011)	Itenov, T.S. et al (2018)	James, M.T. et al (2017)	Lee, B.J. et al (2019)
Model development				
Sample of patients	Patients who survive AKI	Patients admitted to the ICU for at least 24 hours and with AKI	patients with a prehospitalization eGFR of more than 45 $mL/min/1.73m^2$ and who had survived hospitalization with AKI	Adult (age > 18 years) who developed dialysis-requiring AKI (AKI-D)
Study design	Prospective cohort study	Prospective cohort study	Prospective cohort study	Retrospective cohort study
Number of centers	1 center	9 academic ICUs	Multicenter (population-based repository)	21 hospitals
AKI definition	RIFLE	KDIGO	KDIGO	RRT + SCr > 50% rise
Derivation cohort sample size	5351	568	9973	2214
Derivation time period	October 1999 - December 2005	2006 - 2010	April 2004 - March 2014, with follow-up to March 2015	January 2009 - September 2015
The outcome of interest	Risk for progression to CKD stage 4	Recovery after AKI within 28 days	Progression of AKI to advanced CKD	Recovery after dialysis-requiring AKI within 90 days
Number of prediction models	Three logistic regression models	Two cause-specific Cox regression models: one for the hazard of recovery and one for death without recovery	Five multivariate logistic regression	Two models: Logistic regression and classification and regression tree (CART)
Predictor selection method (e.g.full model approach, backward elimination)	Model1: stepwise logistic regression, Model2: based on the most heavily weighted factors from the model1, Model3: based on sentinel clinical events	Model1: most likely predictors, Model2: full model	Stepwise backward logistic regression at $P < 0.05$ with bootstrap selection (1000 samples)	Stepwise logistic regression with bootstrap selection (1000 samples)
Incidence of outcome	13.6% entered CKD4	15.1% risk of not recovering	2.7% developed advanced CKD	59.1% not recovered after AKI-D
Validation method				
Validation cohort sample (eg. split sample, bootstrap)	Separate cohort	Separate cohort	Internal (one-third of derivation cohort) and separate cohort	Internal validation (10-fold cross-validation)
Validation cohort sample size	11589	766	2761 (external cohort)	-
Validation time period	October 1999 - December 2005	1 January 2012 - 31 December 2013	June 2004 - March 2012, with follow-up to March 2013	January 2009 - September 2015
Incidence of outcome	8.5% entered CKD4	10% risk of not recovering	2.2% developed advanced CKD	59.1% not recovered after AKI-D
Performance statistics	$c-statistics = 0.81\tilde{}0.82$	AUROC= 73.1% for predicting recovery	$c-statistic = 0.87$	Logistic regression: $c-index = 0.645$, CART: $c-index = 0.61$
Model performance statistics: calibration	Not reported	Calibration plot used, noted as nicely calibrated	$P(slope) = 0.92, 0.88, 0.8, 0.89, 0.67$	Calibration plot used, noted as excellent calibration

Table 5.3 continued:: AKI-outcome prediction models.

	Chen, Z. et al (2021)	He, J. et al (2021)	Pike, F. et al (2015)	Huang, C.Y. et al (2022)
Model development				
Sample of patients	Patients diagnosed with cardiac surgery-associated AKI (CSA-AKI)	Patients with sepsis-associated AKI	Critically ill patients receiving RRT with AKI	ICU patients with AKI-3
Study design	Prospective cohort study	Prospective cohort study	Prospective cohort study	Prospective cohort study
Number of centers	1 center	1 center	Multicenter	Multicenter (seven ICUs)
AKI definition	Not mentioned	KDIGO	Not mentioned	KDIGO
Derivation cohort sample size	196	209	1124	229
Derivation time period	not mentioned	January 2015 - December 2020	November 2003 - July 2007	August 2007 - November 2010
The outcome of interest	Postoperative AKI requiring RRT or in-hospital death	Predict the occurrence of acute kidney disease (AKD) in patients with sepsis-associated AKI	Renal recovery and mortality for ill patients with AKI requiring RRT at day 60	Two outcomes: 1) complete recovery and 2) complete or partial recovery at hospital discharge
Number of prediction models	Five logistic regression models with different combinations of the 3 selected predictors	Three models: Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM), decision trees, and logistic regression	Four logistic regression models (ATN clinical model, reduced ATN model, LASSO model, stepwise-selected model, and parsimonious model)	Multiple Least absolute shrinkage and selection operator (LASSO) models
Predictor selection method (e.g. full model approach, backward elimination)	LASSO logistic regression and random forests	LASSO	Model1: reduced ATN model, Model2: LASSO, Model3: stepwise logistic regression, Model4: routinely available predictors	Correlation-based feature selection (n=4) and one feature added based on literature
Incidence of outcome	16.3%	55.5%	36.5%	37.55% (completely recovery)
Validation method				
Validation cohort sample (eg. split sample, bootstrap)	Internal validation (bootstrap) and separate cohort	Separate cohort (MIMIC III database)	Internal validation (2-fold split)	Internal validation (stratified 10-fold cross-validation) and separate cohort
Validation cohort sample size	52	509	562	244
Validation time period	Not mentioned	2008 - 2014	November 2003 - July 2007	August 2007 - November 2010
Incidence of outcome	21.1%	46.4%	-	33.20% (completely recovery)
Performance statistics	ROC-AUC=0.971% (0.932-1.000)	Accuracy for LSTM = 97.66%	Renal recovery using model 4: AUROC = 0.76%	Complete recovery: AUROC = 0.53%, complete or partial recovery: AUROC = 0.61%
Model performance statistics: calibration	Calibration score assessed by Brier score and noted as excellent	Calibration plot used, noted as nicely calibrated	HL: P = 0.08–0.45	Calibration plot used

Table 5.4: Quality assessment of model development.

	Chawla et al. [24]	Itenov et al. [85]	James et al. [86]	Lee et al. [104]	Chen et al. [28]	He et al. [66]	Huang et al. [82]	Pike et al. [147]
The baseline study sample is adequately described for key characteristics	Y	Y	Y	Y	Y	Y	Y	Y
Inclusion criteria adequately described	Y	Y	Y	Y	Y	Y	Y	Y
Exclusion criteria are adequately described	Y	Y	Y	Y	Y	Y	Y	Y
Description of the rationale for including predictive variable	N	Y	Y	Y	Y	N	Y	Y
A clear definition or the description is provided for each predictive variable	Y	Y	Y	Y	Y	Y	Y	Y
Continuous variables are reported or appropriate (i.e. not data-dependent) cut points are used	Y	Y	N-goodness-of-fitness tests used to maximize model fit	Y	Y	Y	Standardized to zero, mean, and unit variance	Y
Adequate proportion of the study sample has complete data for prognostic factors	Y	Not reported	Not reported	Y	Not reported	Not reported	Y	Y
Appropriate methods are used if the imputation is used for missing prognostic factor data	Complete case analysis	Not reported	Not reported	Not reported	Not reported	Not reported	Mean for continuous data and the mode for categorical	Multiple imputation
The outcome of interest is clearly defined	Y	Y	Y	Y	Y	Y	Y	Y
The strategy for model building (i.e. inclusion of variables) is appropriate and is based on a conceptual framework or model (i.e. adequate description of mathematical techniques to derive the model)	Y—stepwise multivariate	Y—full model	Y—backward selection	Y—stepwise multivariate	Y—LASSO and random forests	Y—LASSO	Y—correlation-based	Y—LASSO and stepwise multivariate
Internal validation of the model is adequately described (e.g. bootstrapping, cross-validation, or internal validation cohort details are provided)	Not reported	Not reported	Y	Y	Y	Not reported	Y	Y
Model performance statistics were provided with confidence intervals (e.g. ROC curves/c-statistic, HL statistics, likelihood ratios, PPV or NPV)	Y	Y	Y	Y (point estimate only)	Y	Y (point estimate only)	Y (point estimate only)	Y
Sufficient number of events per variable to avoid overfitting (e.g. >10)	Y	Y	Y	Y	Y	N	Y	Y
The model was applied in a cohort independent of the development cohort and the model's predictive performance was assessed	Y	Y	N	N	N	Y	N	Y

Chapter 6

Comparison between cystatin C- and creatinine-based estimated glomerular filtration rate in the follow-up of patients recovering from a stage 3 AKI in ICU

The following chapter has been published in the Journal of Clinical Medicine:

Nateghi Haredasht, F., Viaene, L., Vens, C., Callewaert, N., De Corte, W., & Pottel, H. Comparison between cystatin C- and creatinine-based estimated glomerular filtration rate in the follow-up of patients recovering from a stage 3 AKI in ICU. *Journal of Clinical Medicine* 2022 Dec 7;11(24):7264.

DOI: 10.3390/jcm11247264.

Abstract

Introduction: Acute kidney injury (AKI) in critically ill patients is associated with a significant increase in mortality as well as long-term renal dysfunction and chronic kidney disease (CKD). Serum creatinine (SCr), the most widely used biomarker to evaluate kidney function, does not always accurately predict the glomerular filtration rate (GFR) since it is affected by some non-GFR determinants such as muscle mass and recent meat ingestion. Researchers and clinicians have gained interest in cystatin C (CysC), another biomarker of kidney function. The study objective was to compare GFR estimation using SCr and CysC in detecting CKD over a 1-year follow-up after an AKI-stage 3 event in the ICU, as well as to analyze the association between eGFR (using SCr and CysC) and mortality after the AKI event.

Methods: This prospective observational study used the medical records of ICU patients diagnosed with AKI stage 3. The SCr and CysC were measured twice during the ICU stay and four times following the diagnosis of AKI. The eGFR was calculated using the EKFC equation for SCr and FAS equation for CysC in order to check the prevalence of CKD (defined as eGFR < 60 ml/min/1.73 m^2).

Results: The study enrolled 101 patients, 36.6% of whom were female, with a median age of 74 years (30-92), and a median length of stay of 14.5 days in intensive care. A significant difference was observed in the estimation of GFR when comparing formulas based on SCr and CysC, resulting in large differences in the prediction of CKD. Three months after the AKI event, $eGFR_{CysC} < 25$ mL/min/1.73 m^2 was a predictive factor of mortality later on; however, this is not the case for $eGFR_{SCr}$.

Conclusion: The incidence of CKD was highly discrepant with $eGFR_{CysC}$ versus $eGFR_{SCr}$ during the follow-up period. CysC detects more CKD events compared to SCr in the follow-up phase and $eGFR_{CysC}$ is a predictor for mortality in follow-up but not $eGFR_{SCr}$. Determining the proper marker to estimate GFR in the post-ICU period in AKI stage 3 populations needs further study to improve risk stratification.

6.1 Introduction

Acute Kidney Injury (AKI) is a common clinical syndrome characterized by a rapid decline in kidney function [16, 75, 21, 159]. Two new classification definitions of AKI were proposed in 2004 and 2007, RIFLE and AKIN, respectively. In 2012, the Kidney Disease Improving Global Outcomes (KDIGO) published a clinical guideline to harmonize AKIN and RIFLE diagnostic criteria into one common diagnostic guideline [92]. AKI, particularly AKI stage 3, is associated with a significant increase in mortality, as well as short-term and long-term renal dysfunction, which may ultimately lead to chronic kidney disease (CKD). The prediction of these events (AKI and CKD after AKI) has gained greater attention in recent years. The basis for a good prediction model for AKI and/or CKD is built upon the predictability and the range of values of biomarkers that are being taken into consideration by the model. Currently, serum creatinine (SCr) is the most widely used biomarker to estimate glomerular filtration rate (GFR) which is the best overall index of kidney function [76]. Nevertheless, SCr has some limitations since it depends on muscle mass. Consequently, SCr-based eGFR equations may overestimate the true GFR of critically ill patients, since these patients are suffering from continuous loss of muscle mass [101, 58]. In addition to SCr, cystatin C (CysC) is another biomarker of kidney function which has attracted the attention of researchers and clinicians in recent years. Even though CysC can also be affected by non-GFR determinants, the non-GFR determinants that affect CysC are distinct from those that affect SCr. Smoking status and serum C-reactive protein level are, for instance, independently associated with serum CysC levels [101]. While SCr-based eGFR-equations are widely used, several CysC-based eGFR-equations have been validated [83, 149, 58]. The most commonly used equations to calculate eGFR include the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation [110], and the full age spectrum (FAS) equation [149, 150]. Recently, Pottel et al., introduced an optimized FAS SCr-based equation named the European Kidney Function Consortium (EKFC) equation, to estimate GFR [148]. Predicting CKD following AKI is a highly important yet missing topic in the AKI research field. The earlier CKD is diagnosed after AKI, the less intensive utilization of resources and the better the prevention of morbidity and mortality [90]. Delanaye et.al., [39] compared the performance of CysC and SCr as biomarkers for estimating GFR in 47 critically ill patients and concluded that CysC significantly outperformed SCr for the detection of an impaired GFR. Moreover, a recent study has been conducted on 22,488 critically ill patients to compare long-term mortality risk prediction by eGFR using an SCr-based equation (CKD-EPI), a CysC-based equation (CAPA) [58], and a composite SCr/CysC-based equation (CKD-EPI). Results showed that the single biomarker CysC equation performed better compared to the SCr or composite equations when estimating GFR for risk prediction purposes

in critically ill patients [67]. Moreover, Gharaibeh et al., claimed that CysC decreases before SCr in most hospitalized patients with acute kidney injury and therefore predicts renal recovery earlier than SCr [55]. Despite evidence confirming the superiority of CysC over SCr in detecting AKI in the intensive care unit (ICU), there is no similar evidence comparing CysC to SCr in detecting CKD in the post- ICU period after experiencing AKI in ICU. Researchers have mostly used SCr to estimate GFR in studies that attempted to detect renal recovery or CKD in survivors of AKI after discharge. To our knowledge, only one study by Rimes-Stigare et al., [156] focused on the occurrence of CKD and Acute Kidney Disease (AKD) in AKI survivors three months after AKI that used both SCr and CysC. Their result showed significant renal impairment of at least 25% according to SCr-based CKD and 67% when classified using CysC-estimated GFR.

In this study, we investigated whether the eGFR based on CysC differed from the eGFR based on SCr during the ICU stay and the follow-up period in adult ICU patients who had experienced an AKI stage-3 event. Additionally, we examined whether eGFR determined by using the CysC-based FAS equation would further improve the correlation between eGFR and CKD diagnosis, as well as adjusted risks of death in the follow-up, compared to the use of SCr-based eGFR. We hypothesized that CysC might provide additional benefits in post-ICU and be better related to the outcomes (CKD and mortality) considering the inherent risk of muscle wasting that affects SCr.

6.2 Materials and methods

6.2.1 Study design and participants

This study is a prospective observational study where we used medical records of ICU patients aged > 18 years who are diagnosed with AKI stage 3 during their ICU stay in AZ Groeninge hospital in Kortrijk, Belgium, between September 2018, and October 2020. Exclusion criteria were patients with a baseline eGFR < 30 ml/min/1.73 m^2 estimated by CKD-EPI [110], patients with renal replacement therapy (RRT) initiated before admission to the ICU, patients with a kidney transplant, patients with therapy restrictions with shift to palliative care, and patients who received extracorporeal blood purification techniques for reasons other than AKI. Demographic data, comorbidity data, the severity of illness scores (APACHE 2), admission diagnosis, laboratory data, and data concerning kidney function (serial SCr measurements, oliguria, the time when AKI stage 3 developed, urinary analysis, etc.) were reported during the ICU stay. These data were augmented with regular GFR estimation by the CKD-EPI formula using both SCr and CysC biomarkers. SCr measurements as in clinical practice in ICU were collected every morning; however, CysC measurements

were not part of routine clinical practice and were obtained only at the time of admission to ICU and at the time of diagnosis of AKI (in most patients with a limited time lag). Furthermore, these patients have been followed up at the nephrology department at 3, 6, 9, and 12 months after AKI stage 3 diagnosis in ICU. During these follow-up visits, the eGFR again was determined using both biomarkers.

6.2.2 Definitions - Acute Kidney Injury criteria and calculations

KDIGO criteria for AKI stage 3 have been used for the inclusion of patients based on SCr or urine output (UO). KDIGO defines stage 3 as an increase in SCr up to 3 times from baseline within a 7-day period or UO < 0.3 ml/kg/h for ≥ 24 hours [92]. In this study, true baseline SCr was available for patients who had an SCr measurement from an earlier visit (previously to their hospital or ICU admission). In the absence of such records, baseline SCr was considered the first record of a patient's hospitalization prior to being admitted to the ICU.

6.2.3 Serum creatinine and cystatin C measurement

All SCr measurements were performed with an Enzymatic method that is traceable to the isotope dilution mass spectrometric method (IDMS), which is the internationally approved reference method for measuring creatinine. In addition, CysC concentrations were measured by Liège University Hospital using a particle-enhanced nephelometric immunoassay on the BNII nephelometer (Siemens Healthcare Diagnostics, Marburg, Germany). The assay was calibrated against the international certified reference material ERM-DA471/IFCC for CysC.

6.2.4 Evaluation of glomerular filtration rate

The SCr-based estimated glomerular filtration rate ($\mathrm{eGFR_{SCr}}$) was calculated according to the EKFC equation introduced by Pottel et al., in 2020 [148]:

$$EKFC - eGFR = \begin{cases} 107.3 \times (SCr/Q_{crea})^{-0.322} [\times (0.990)^{(Age-40)} \; if \; age > 40 \; years], & SCr/Q_{crea} < 1. \\ 107.3 \times (SCr/Q_{crea})^{-1.132} [\times (0.990)^{(Age-40)} \; if \; age > 40 \; years], & SCr/Q_{crea} \geq 1. \end{cases} \quad (6.1)$$

The EKFC equation is based on normalized SCr (SCr/Q_{crea}) where Q_{crea} is the median SCr from healthy populations, which is 0.70 mg/dL for females and 0.90 mg/dL for males.

The CysC-based estimated glomerular filtration rate ($\mathrm{eGFR_{CysC}}$) was calculated according to the full age spectrum (FAS) equation introduced by Pottel et al., in 2017 [150, 149]:

$$FAS - eGFR = \begin{cases} 107.3/(CysC/Q_{Cys}), & for\ 2 \leq age \leq 40\ years. \\ 107.3/(CysC/Q_{Cys})[\times(0.988)^{(Age-40)}], & for\ age > 40\ years. \end{cases} \quad (6.2)$$

The FAS equation is based on normalized CysC ($CysC/Q_{Cys}$) where Q_{Cys} is the median CysC from healthy populations, which is 0.82 mg/L when age < 70 years and 0.95 otherwise, both for males and females.

6.2.5 Outcomes

The primary outcome was the post-ICU incidence of CKD after experiencing AKI stage 3, based on decreased eGFR by SCr and CysC levels. To evaluate whether post-ICU eGFR values measured by each marker were clinically valid, we compared associations of eGFR, detected by CysC versus SCr level, with mortality after ICU discharge as a clinical endpoint.

6.2.6 Statistical analysis

Continuous variables were presented as medians with interquartile ranges (IQR) and categorical variables were expressed as percentages. Correlation between all measurements of the biomarkers was assessed using Pearson and Spearman's correlation coefficient. The normality of the distributions was assessed with the Shapiro–Wilk test. A Mann–Whitney U test / Wilcoxon rank-sum test was used to compare continuous variables of independent subgroups. The associations of eGFR_{SCr} and $\text{eGFR}_{\text{CysC}}$ with mortality were analyzed using Cox proportional hazard regression and logistic regression models, adjusted for covariates age, sex, length of stay (LoS), and dialysis in ICU. Kaplan-Meier survival curves were plotted for SCr or CysC-based eGFR < 25 and eGFR ≥ 25 mL/min/1.73 m^2 in the first follow-up measurement and compared using the log-rank test. Given the multiple visits per patient during follow-up, we used linear mixed models, derived slopes and intercepts for both eGFR_{SCr} and $\text{eGFR}_{\text{CysC}}$ (only for the follow-up period), and compared the slopes and intercepts. For mixed-effect models, subjects and time (days in follow-up) are treated as random effects. Supplementary materials provide more details about the statistical analysis. A two-tailed p-value of $p < 0.05$ was considered statistically significant. Analyses were carried out using R Statistical Software (version 4.0.5).

6.3 Results

6.3.1 Patients

A total of 101 critically ill patients (37 females, median (IQR) age of 74 (30–92) years) who developed AKI stage 3 were included in this study. Characteristics of patients on ICU admission and after discharge are shown in 6.1. Patients who survived ICU and who were followed up successfully with no dropout had six different measurements of both SCr and CysC: at the time of admission to ICU, at the time of developing AKI stage 3, and four follow-up times (every three months up to one year after AKI diagnosis). 45% of the cohort (n=46) patients received dialysis with a median of 13 (1-160) days during ICU stay and the mortality rate during the study was 42.6% (n=43). 24 patients died during ICU stay and 19 patients died in the follow-up phase, of which 2 died between ICU discharge and the first follow-up. The number of patients attending follow-up visits decreased due to patient dropouts and mortality. Table 6.2 summarizes the median days after hospital discharge together with the number of SCr, CysC, and both SCr/CysC measurements in each follow-up visit. In follow-up visits, the number of patients with SCr and CysC measurements may not match due to storage and transport issues.

6.3.2 Correlation between serum creatinine and cystatin C

Figure 6.1 shows the Spearman correlation between SCr and CysC, before any adjustment (for age and/or sex), during ICU stay, and during the follow-up, phase using all measurements. As shown in Figure 1, SCr and CysC are positively related, but CysC levels off at 8 mg/L while SCr can go higher than 15 mg/dL. Since SCr and CysC tend to move in the same direction, but not necessarily at the same rate, their relationships are monotonic. Due to the monotonic relationship between the variables, we chose the Spearman correlation coefficient in Figure 6.1. In Figure 6.1, graphs A and B show the relation between SCr and CysC during ICU stay for males, and females, respectively, and graphs C and D show the relation during the follow-up phase for males, and females, respectively. Figure 1 illustrates the relatively high correlation between the two biomarkers during ICU; however, the correlation during the follow-up phase is much lower, specifically for males. Note that the axis values are different during ICU and the follow-up. Rescaling SCr to SCr/Q and CysC to CysC/Q did not change the correlation coefficients.

Table 6.1: Patient characteristics.

Characteristics	**All patients (N=101)**
Demographics	
Female sex, n (%)	37 (36.6%)
Age, years	74 (30-92)
Body weight, kg	83 (45-150)
Body mass index, kg/m^2	27.7 (17-57)
ICU types	
MICU, n (%)	81 (80%)
SICU, n (%)	17 (16.8%)
Trauma, n (%)	3 (2.9%)
Admission diagnosis (%)	
Community-acquired pneumonia (CAP)	10%
Cardiac disease	9%
Acute respiratory failure	8%
Sepsis	7%
Aspiration pneumonia	6%
Pulmonary edema	4%
Other diagnoses	56%
Results	
Length of stay in ICU, days	14.5 (1-160)
Cystatin C (mg/L) at ICU admission	1.94 (0.67-8.06)
Creatinine (mg/dL) at ICU admission	1.98 (0.31-12.64)

Values are median (IQR) or n(%).

Table 6.2: Patients' follow-up information after hospital discharge.

	1^{st} follow-up	2^{nd} follow-up	3^{rd} follow-up	4^{th} follow-up
Number of survivors for follow-up	75	61	48	40
Number of dropouts	7	7	3	5
Median follow-up days	37	142	229	337
Number of patients with SCr values	68	54	45	35
Number of patients with CysC values	62	39	34	26
Number of patients with SCr and CysC	60	39	34	25

6.3.3 Evaluation of eGFR using serum creatinine and cystatin C

Results of $eGFR_{SCr}$ against $eGFR_{CysC}$ during ICU stay, and follow-up period is presented in Figure 6.2 plots A and B. Pearson's correlation test between the two eGFRs during the ICU stay (plot A) is 0.82; however, during the follow-up phase (plot B) the correlation decreased to 0.7. We have used Pearson's test since both equations are supposed to predict the same value (GFR); hence, they should be linearly related ideally with a correlation coefficient of '1' and slope = 1 and

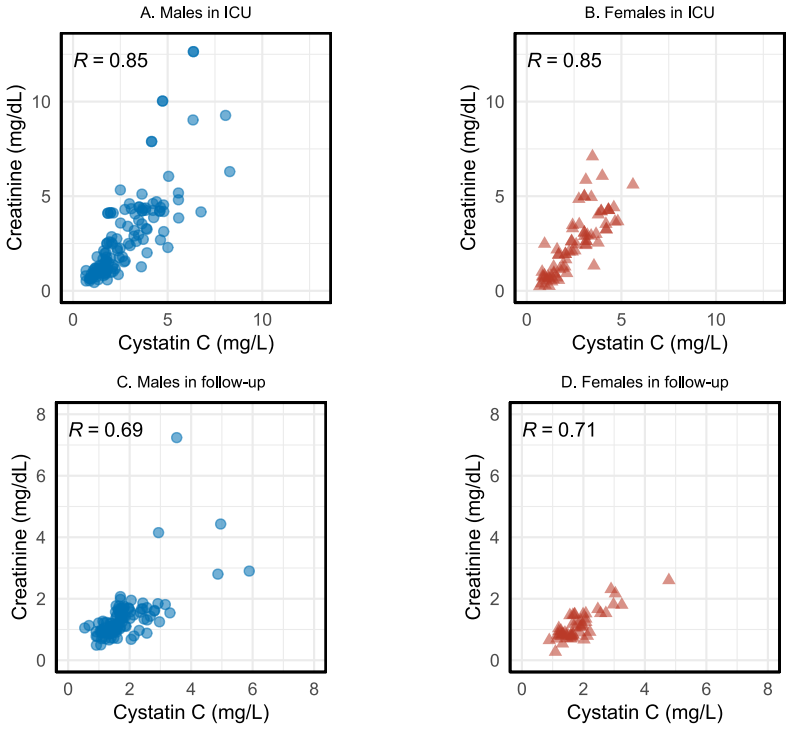


Figure 6.1: Spearman correlation coefficient R for males and females for SCr and CysC in the ICU stay and follow-up phase.

intercept = 0. However, our results demonstrate systematic deviation from the identity line, showing that two biomarkers behave significantly differently in the follow-up phase ($p - value < 0.0001$; Wilcoxon test). No significant deviation from the identity line has been observed during the ICU stay (p-value=0.1; Wilcoxon test). Also, the Bland-Altman analysis for the ICU stay and follow-up phase is presented in Table Figure A.6 in the Appendix A. We see that during the ICU stay, the average difference between eGFR_{SCr} and $\text{eGFR}_{\text{CysC}}$ is near zero; however, it is nearly 15 in the follow-up phase.

Figure 6.3 shows boxplots for all measurements of eGFR based on SCr and CysC during ICU stay (at admission and AKI event time) and during each follow-up visit. Results show that eGFR_{SCr} levels were higher during each follow-up visit compared to $\text{eGFR}_{\text{CysC}}$. Note that the number of patients differs in each follow-up visit due to drop-out or death. Table A.6 in the Appendix A provides data for the median eGFR value and interquartile range (IQR) for patients with both biomarkers measured. There is no statistically significant difference

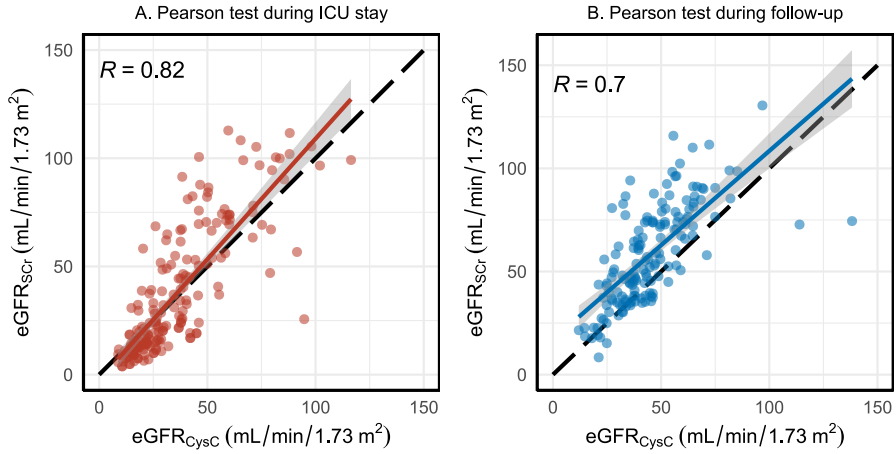


Figure 6.2: eGFR (mL/min/1.73 m^2) comparison using SCr and CysC in ICU stay and follow-up phase. The red and blue curves are fitted linear regression models in ICU and follow-up, respectively, and the faded zones are the confidence intervals around the lines. The black dashed line shows the identity line.

between the two eGFR values during the ICU stay (Wilcoxon signed-rank test); however, this difference is significant during each follow-up visit (Wilcoxon signed-rank test) except for the last follow-up (4^{th}) which is probably due to the small sample size.

The within-subject evolution of the eGFR for alive patients using both SCr and CysC from the first measurements in ICU until the last follow-up is shown in Figure 6.4. We have used the measurements of only alive patients since only the survivors could have data at the latest visits (some curves end before the follow-up moment due to dropouts).

As shown in Figure 6.4, $\text{eGFR}_{\text{CysC}}$ increases steadily from the time the AKI is diagnosed (time point 2) onward, whereas eGFR_{SCr} plateaus at the first follow-up (time point f1), which may reflect the fact that kidney function improves (thus decreasing SCr) and muscle mass increases when patients are recovering

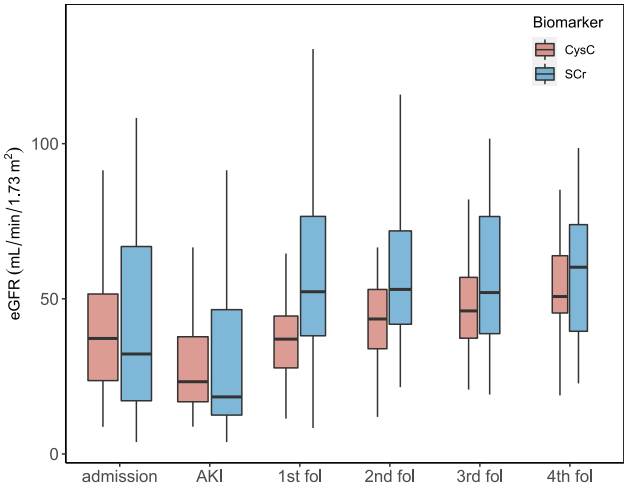


Figure 6.3: Comparison of eGFR SCr and eGFR CysC during ICU stay and each follow-up phase.

from their ICU stay (thereby increasing SCr). The combined effect may be that SCr does not change much; however, CysC reflects true renal recovery. Figure A.7 in the supplementary material shows the individual weight evolution during the follow-up period. Results show that the patients start gaining weight after the second follow-up. The mixed model analysis (Table 6.3) confirms the LOESS ([29]) observations (Figure 6.4) during follow-up: eGFR_{SCr} starts at a much higher average level but shows no change over time during the follow-up period, while $\text{eGFR}_{\text{CysC}}$ is lower at month 3 (first FU) but shows a significant increase during follow-up. Moreover, the intercepts are significantly different since the 95% confidence intervals (CI) do not overlap. Table 6.3 supports our hypothesis regarding the effect of muscle mass on SCr. Considering that the intercept for eGFR_{SCr} in the mixed-effect model is much higher than the intercept for $\text{eGFR}_{\text{CysC}}$ and furthermore that the slope for eGFR_{SCr} is not changing while the slope for $\text{eGFR}_{\text{CysC}}$ is increasing (regaining kidney function) confirms that according to the SCr, it seems as if the kidneys have already recovered at the first follow-up, while CysC-based eGFR still shows ongoing recovery during follow-up.

Table 6.4 indicates the number of patients with eGFR < 60 mL/min/1.73 m^2 and eGFR ≥ 60 mL/min/1.73 m^2 based on SCr and CysC in each follow-up visit. We see large differences in the incidence of chronic kidney disease (defined as eGFR < 60 ml/min/1.73 m^2) using the two biomarkers during follow-up visits. Specifically, in the first follow-up, as shown in Table A.7 in Supplementary

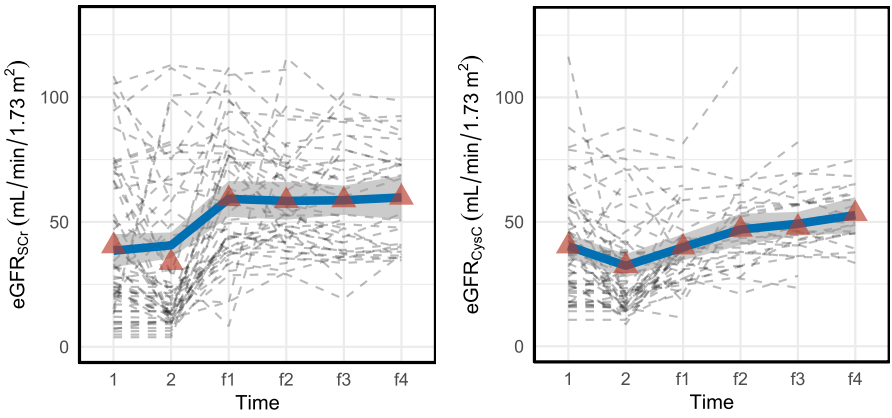


Figure 6.4: Within-subject evolution of eGFR for alive patients from the first day in ICU until the last follow-up. The dashed gray lines represent each subject, the red triangles show the average eGFR values at that specific time point, and the blue lines are smooth curves obtained via LOESS. The gray band is a 95% confidence band for the regression line

Table 6.3: The output of the mixed effect model results.

	eGFR$_{SCr}$			eGFR$_{CysC}$		
Predictors	Estimates	CI	p	Estimates	CI	p
Intercept	57.25	50.50–64.00	<0.001	37.85	33.8–42.27	<0.001
Time	0.003	-0.01–0.02	0.7	0.041	0.01–0.07	0.004

material, 19 patients were classified as having CKD using eGFR$_{CysC}$; however, eGFR$_{SCr}$ had classified them as having no CKD. The difference between the two biomarkers in detecting CKD in the first follow-up was statistically significant (McNemar's chi-squared = 14.45, p-value=0.000143).

Table 6.4: Number of patients with eGFR $<$ 60 mL/min/1.73 m^2 and eGFR $\geq$ 60 mL/min/1.73 m^2 based on SCr and CysC in each follow-up visit.

	$eGFR_{SCr}$		$eGFR_{CysC}$	
	<60	>60	<60	>60
Visit 1 (n=60)	34	26	52	8
Visit 2 (n=39)	23	16	34	5
Visit 3 (n=34)	20	14	29	5
Visit 4 (n=25)	10	15	15	10

6.3.4 The associations between eGFR and outcome

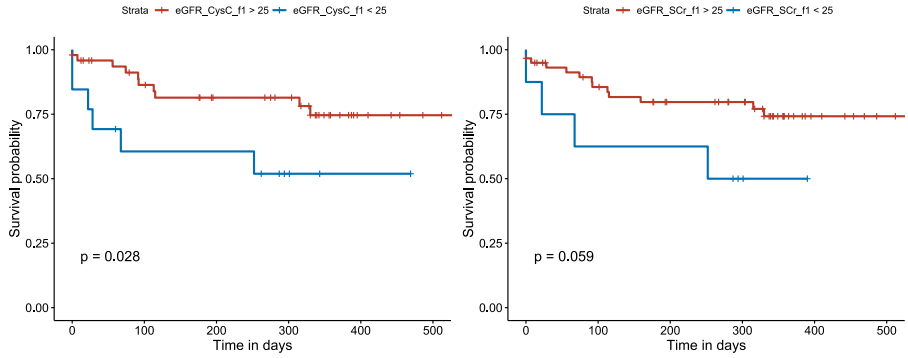
We also evaluated whether there were differences between those with CKD and those without CKD based on either eGFR SCr or CysC levels during each follow-up time (Tables A.8, A.9, A.10, and A.11 in Appendix A). Stages of CKD are defined using the KDIGO guidelines. The results of each follow-up show a large difference in the patients' classification of CKD using SCr and CysC. For instance, according to Table A.8 in Appendix A, $eGFR_{SCr}$ classifies the majority of patients (n=19) as GFR category 2, on the other hand for $eGFR_{CysC}$, the majority (n=29) belong to class CKD3B (moderate to severely decreased).

There were 43 (42.6%) deaths during the study, of which 24 occurred during the ICU stay and 19 during the follow-up period. Univariate Cox proportional hazard regression models were performed to examine the risk factors associated with mortality in ICU with all patients included and mortality in follow-up with patients who survived ICU (Table 6.5).

The analysis of the association between eGFR and mortality in ICU has been done on the whole population (n=101), and we have considered variables, age, gender, average $eGFR_{CysC}$ in ICU, and average $eGFR_{SCr}$ in ICU. The average $eGFR_{CysC/SCr}$ in ICU is the average of over two eGFRs using SCr and CysC during ICU stay.

In analyses of the association between eGFR and mortality in the follow-up phase, patients who survived the ICU and appeared at the first follow-up visit were included. Variables including age, gender, length of stay in ICU (LoS ICU), dialysis in ICU, and reduced $eGFR_{CysC}$ and $eGFR_{SCr}$ in the first follow-up were considered in the model. Reduced $eGFR_{CysC/SCr}$ in the first follow-up was defined as eGFR $<$ 25 mL/min/1.73 m^2.

Results of Cox proportional hazard regression models in Table 6.5 demonstrate that age was a significant risk factor for mortality in ICU and that a patient who has eGFR based on CysC below 25 ml/min/1.73 m^2 at 1_{st} follow-up, has a significantly increased risk for mortality compared to a patient who has eGFR



(a) eGFR levels using CysC

(b) eGFR levels using SCr

Figure 6.5: Kaplan–Meier survival curves according to eGFR levels using CysC (left) and SCr (right) in patients with eGFR below and above 25 ml/min/1.73 m^2 in the first follow-up measurement. There are 49 deceased patients with eGFR CysC below 25 and 13 with eGFR CysC above 25, 60 deceased patients with eGFR SCr below 25, and 8 with eGFR SCr above 25.

based on SCr below 25 ml/min/1.73 m^2. Kaplan-Meier curves using eGFR $<$ 25 and $\geq$ 25 as strata illustrate these findings (Figure 6.5). We have divided the patients who survived during ICU, into two groups based on eGFR value at 1^{st} follow-up using both biomarkers to investigate whether having eGFR below or above 25 ml/min/1.73 m^2 is predictive of mortality during follow-up (Figure 6.5).

Table 6.5: Univariate Cox regression models for mortality in ICU and follow-up.

Variable	Mortality in ICU		Mortality in follow-up	
	HR	95% CI	HR	95% CI
Age	1.05*	1.01-1.09	1.03	0.99-1.07
Gender (male)	0.93	0.41-2.13	2.022	0.67-6.10
LoS in ICU	-	-	0.98	0.96-1.01
Dialysis in ICU	-	-	1.86	0.87-3.95
Reduced $eGFR_{CysC}$ in the first follow-up	-	-	3.32*	1.2-9.2
Reduced $eGFR_{SCr}$ in the first follow-up	-	-	2.8	0.91-8.65
Average $eGFR_{CysC}$ in ICU	1	0.97-1.01	1.00	0.97-1.02
Average $eGFR_{SCr}$ in ICU	0.98	0.96-1.01	1.00	0.98-1.02

6.4 Discussion

AKI is very common in the critically ill. Spontaneous resolution (or rapid response to treatment) occurs in some patients, even after experiencing the most

severe AKI Stage 3 event. Despite its relative non-specificity, SCr remains the gold standard for defining AKI and for follow-up after an AKI event. However, less is known about CysC during the follow-up phase after experiencing an AKI episode. The present study investigates whether the use of the CysC has advantages over SCr as a biomarker for renal function for adult ICU patients who had experienced such an AKI stage 3 event. First, by comparing the evolution of the two biomarkers and estimated GFR during the ICU stay and post-AKI follow-up, we discovered that they behave differently after the ICU discharge, and the correlation between the two GFR estimates drops during the follow-up period. Several articles suggest that SCr may result in an overestimation of recovery by ignoring the decrease in SCr due to the loss of muscle mass that occurs during critical illness [51, 152]. The majority of our AKI stage 3 patients (almost 30%) developed this event on the day of admission to the ICU or the day following admission; therefore, SCr was not affected by the loss of muscle mass during the ICU stay, resulting in a higher correlation between SCr-based and CysC-based eGFR during the stay.

Using different statistical analyses, we compared creatinine- and cystatin C-based estimates of GFR during ICU stay and the follow-up period. On admission, we observed that both eGFRs are approximately the same, which strongly supports the hypothesis that the loss of muscle mass explains the differences observed over time in the follow-up phase. In the follow-up period, we found a significant difference in eGFR values between the two biomarkers. In particular, we saw $\text{eGFR}_{\text{CysC}}$ increase steadily from the time the AKI is diagnosed onward, whereas eGFR_{SCR} plateaus at the first follow-up which may reflect the fact that kidney function improves (thus decreasing SCr) and muscle mass increases when patients are recovering from their ICU stay (thereby increasing SCr). The combined effect may be that SCr does not change much. Due to the fact that CysC is not affected by muscle mass, there is no double effect present for CysC

Moreover, the occurrence of low $\text{eGFR}_{\text{CysC}}$ in the follow-up after AKI was more frequent than the occurrence of low eGFR_{SCr}, especially during the first visit, about 1 month after hospital discharge. Two phenomena might explain the high eGFR_{SCr} values in the first follow-up. First, patients may not yet be recovered from ICU stay, leading to a lower SCr (higher eGFR_{SCr}). Second, patients recovering from AKI have their kidney function improving, which consequently leads to a lower SCr value (higher eGFR_{SCr}). It was suggested in [51] that although follow-up care pathways should be tailored to individual conditions, reassessment of renal function 90 days after discharge from the hospital is more reasonable, in order to allow time for the recovery of muscle mass as well as any further improvement of renal function. Although no significant increase from the first follow-up, our results also confirm that the patients start gaining weight after the 2^{nd} follow-up. In addition to the effect of diet and muscle mass

on creatinine production, overestimation of kidney function in AKI patients due to the elimination of creatinine by tubular secretion could explain these differences in $eGFR_{SCr}$ and $eGFR_{CysC}$ [168, 27].

Furthermore, according to our results, CKD incidence was far higher when GFR was estimated using CysC than with SCr, which was a confirmation of the findings in the study by Rimes-Stigare et al., [156]. We observed that $eGFR_{SCr}$ tends to classify more patients towards the less severe stages compared to $eGFR_{CysC}$. Both biomarkers give similar GFR estimates at steady-state, providing an acceptable correlation with measured GFR [189, 182]. It is worth mentioning that we only considered the measurements in which both markers were measured. Our results suggest that the patients may be classified differently according to the biomarker used. Differences in the incidence of CKD by the two biomarkers seen in our study could be related to a number of factors, including the loss of skeletal muscle mass and strength that occurs during an ICU stay and affects SCr levels even after discharge. It might also be due to the different abilities of CysC-based equations and SCr-based equations to estimate measured GFR in different populations like elderly patients. Study results on the elderly have shown that when SCr and CysC are combined, GFR estimates are more accurate and precise [206], while SCr-based equations are most inaccurate [45].

Since the surveillance of all patients would be expensive and impractical, we must establish how to determine renal function during post-AKI follow-up best. As in our results, we saw that a patient who has eGFR based on CysC below 25 ml/min/1.73 m^2 at 1_{st} follow-up, has a significantly increased risk for mortality compared to a patient who has eGFR based on SCr below 25 ml/min/1.73 m^2; hence, clinicians should look at $eGFR_{CysC}$ instead of $eGFR_{SCr}$ at the first follow-up. After evaluating different cut-offs, 25 was chosen because it gave us the best 'survival' discrimination between the two eGFR-equations. Other cut-offs failed to reach significance probably because there were few participants. Future studies should validate this cut-off.

Even though our findings are intriguing and may be clinically useful, there are some limitations to be considered. First of all, the number of patients in our study was limited and loss of follow-up is present and was due in part to logistical difficulties. Additionally, we did not measure GFR using a gold standard technique since this is not routinely available and is practically impossible in an ICU setting. Moreover, our study was conducted in Europe, and our patient population consisted only of Caucasian patients. Thus, our results cannot be generalized to countries with predominantly black, Asian, or mixed-race populations; moreover, differences in SCr should be taken into consideration due to racial factors. Furthermore, CysC is affected by inflammation and infection; however, we did not adjust CysC for this as CRP was not available in the follow-up.

6.5 Conclusion

Our prospective observational study demonstrated that the incidence of CKD defined with CysC-based eGFR versus SCr-based eGFR during the follow-up period of critically ill patients recovering from an AKI stage 3 in their intensive care unit stay, was highly discrepant. In the follow-up phase, CysC-based eGFR categorized significantly more patients in more severe CKD stages than SCr-based eGFR, and $eGFR_{CysC}$ was a better predictor of mortality, compared to $eGFR_{SCr}$. Accordingly, our study suggested that using SCr alone at follow-up could lead to an underestimate of renal dysfunction (CKD) among AKI stage 3 survivors. Further follow-up is required to evaluate the validity of estimated GFR based on both biomarkers by comparing it to the clinical assessment and progression to dialysis.

Statements

Acknowledgments

The authors would like to thank the patients and staff of AZ Groeninge Hospital without whom this research would not have been possible. The authors also acknowledge the Flemish Government (AI Research Program).

Statement of Ethics

Study approval statement: AZ Groeninge Hospital Ethics Committee and Institutional Review Board approved the study, and the EC study number is AZGS2018070. The study was conducted in accordance with Good Clinical Practice guidelines and the Declaration of Helsinki. **Consent to participate statement:** Written informed consent was obtained by the investigators from the patients or their surrogates before they were enrolled in the study.

Conflict of Interest Statement

The authors declare that they have no competing interests.

Funding Sources

This work was supported by KU Leuven Internal Funds (grant 3M180314).

Author Contributions

Hans Pottel, Wouter De Corte, Celine Vens, and Liesbeth Viaene designed the study. Nico Callewaert, Liesbeth Viaene, and Wouter De Corte collected the

dataset. Fateme Nateghi Haredasht performed the statistical analysis and wrote the first draft of the manuscript with input from Hans Pottel, Wouter De Corte, Celine Vens, and Liesbeth Viaene. The manuscript was reviewed and approved by all authors before submission.

Data Availability Statement

The data supporting the results of this study cannot be made publicly available due to the lack of approval from our ethics committee in this regard.

Chapter 7

Predicting survival outcomes in the presence of unlabeled data

The following chapter has been published in Machine Learning:

Nateghi Haredasht, F., & Vens, C. Predicting survival outcomes in the presence of unlabeled data. *Machine Learning.* 2022 Nov;111(11):4139-57.

DOI: 10.1007/s10994-022-06257-x.

Abstract

Many clinical studies require the follow-up of patients over time. This is challenging: apart from frequently observed drop-out, there are often also organizational and financial challenges, which can lead to reduced data collection and, in turn, can complicate subsequent analyses. In contrast, there is often plenty of baseline data available of patients with similar characteristics and background information, e.g. from patients that did not consent to be followed over time or from patients that fall outside the study time window. In this article, we investigate whether we can benefit from the inclusion of such unlabeled data instances to predict accurate survival times. In other words, we introduce a third level of supervision in the context of survival analysis, apart from fully observed and censored instances, we also include unlabeled instances. We propose three approaches to deal with this novel setting and provide an empirical comparison over fifteen real-life clinical and gene expression survival datasets. Our results demonstrate that all approaches are able to increase the predictive performance over independent test data. We also show that integrating the partial supervision provided by censored data in a semi-supervised wrapper approach generally provides the best results, often achieving high improvements, compared to not using unlabeled data.

7.1 Introduction

Many clinical studies require following subjects over time and measuring the time until a certain event is experienced (e.g., death, progression, hospital discharge, etc). The resulting collected datasets are typically analyzed with survival analysis techniques. Survival analysis is a branch of statistics that analyzes the expected duration until an event of interest occurs [99]. Censoring is an important concept in survival analysis which makes it challenging compared to other analytical methods. Censoring can occur due to various reasons, such as drop-out, and means that the observed time is different from the actual event time. In the case of right censoring, for instance, we know that the actual event time is greater than the observed time [72].

Traditional survival analysis methods include the Cox Proportional Hazards model (CPH) [36]. CPH is basically a linear regression model that predicts simultaneously the effect of several risk factors on survival time. However, these standard survival models encounter some challenges when it comes to real-world datasets. For instance, they cannot easily capture nonlinear relationships between the covariates. In addition, in many applications, the presence of high-dimensional data is quite common, e.g., gene expression data; however, these traditional methods are not able to efficiently deal with such high-dimensional data. As a result, machine learning-based techniques have become increasingly popular in the survival analysis context over recent years [202]. Applying machine learning methods directly to censored data is challenging since the value of a measurement or observation is only partially known. Several studies have successfully modified machine learning algorithms to make use of censored information in survival analysis, e.g., decision trees [56], artificial neural networks (ANN) [46], and support vector machines (SVM) [96] to name a few. Popular ensemble-based frameworks include bagging survival trees [77] and random survival forests [84]. Also, more advanced learning tasks such as active learning [198] and transfer learning [114] have been extended toward survival analysis.

Long-term follow-up of patients is often expensive, both time- and effort-wise and financially. As a result, the number of subjects that are included in a study and followed in time is often limited, although many more subjects may exist (e.g., through retrospective data collection) that meet the inclusion/exclusion criteria of the follow-up study. If the study aims to predict outcomes based on variables collected at baseline, then we hypothesize that these extra (unlabeled) data points might actually boost the predictive performance of the resulting model, if used wisely. This corresponds to a semi-supervised learning set-up [23], which deals with scenarios where only a small part of the instances in the training data have an outcome label attached, but the rest is unlabeled. To our knowledge, such a semi-supervised learning set-up has never been investigated in the context of survival analysis, and with this article, we aim to fill this gap.

Including unlabeled instances in a survival analysis task leads to three distinct subsets of data, that differ in the amount of supervised information they contain: a set of (1) fully observed, (2) partially observed (censored), and (3) unobserved data points. Our goal is to look at these three subsets of data altogether. In particular, we address two research questions: (1) can the predictive performance over an independent test set be increased by including unlabeled instances (i.e., does the semi-supervised learning setting carry over to the survival analysis context)?, and (2) what is the best approach to integrate the 3 subsets of data in the analysis? To address this second question, we propose and compare three different approaches. The first approach is to treat the unlabeled instances as censored with the censoring time equal to zero and apply a machine learning-based survival analysis technique. For the second approach, we apply a standard semi-supervised learning approach. In particular, we use the widely used self-training wrapper technique [209, 113]. This technique first builds a classifier over the labeled (in our case, observed and censored) data points and iteratively augments the labeled set with highly confident predictions over the unlabeled dataset. In the third approach, we propose an adaptation of the second one, in which we initially add the censored instances to the unlabeled set, and exploit the censored information in the data augmentation process, to decide how many instances to add to the labeled set in each iteration. In all three approaches, we use random survival forests as base learner [84]. In order to answer the research questions, we apply and compare the approaches using fifteen real-life datasets from the healthcare domain.

The outline of the chapter is as follows. Section 7.2 introduces the background and reviews some concepts of the employed models including random survival forest and self-training approaches. Section 7.3 describes related work. In section 7.4, three proposed approaches are introduced, two of which are a self-training-based framework that copes with the survival data. Section 7.5 presents the experimental set-up, including dataset description, unlabeled data generation, performance evaluation, and comparison methods, and parameter instantiation. Results are presented in section 7.6. Conclusions are drawn in section 7.7.

7.2 Background

In this section, we first review some concepts of using machine learning methods for survival analysis. Afterward, we explain the self-training technique and how one can apply it to a survival analysis problem.

7.2.1 Random survival forest

Random survival forests are well-known ensemble-based learning models that have been widely used in many survival applications and have been shown to be superior to traditional survival models [125]. Random survival forest (RSF) [84] is quite close to the original Random Forest by Breiman [17]. The random forest algorithm makes a prediction based on tree-structured models. Similar to the random forest, RSF combines bootstrapping, tree building, and prediction aggregating. However, in the splitting criterion to grow a tree and in the predictions returned in the leaf nodes, RSF explicitly considers survival time and censoring information. RSF has three main steps. As the first step, it draws B bootstrap samples from the original data. In the second step, for each bootstrap sample, a survival tree is grown. At each node of a tree, p candidate variables are randomly selected, where p is a parameter, often defined as a proportion of the original number of variables. The task is to split the node into two child nodes using the best candidate variable and split point, as determined by the log-rank test [166]. The best split is the one that maximizes survival differences between the two child nodes. Growing the obtained tree structure is continued until a stop criterion holds (e.g., until the number of observed instances in the terminal nodes drops below a specified value). In the last step, the cumulative hazard function (CHF) associated with each terminal node in a tree is calculated by the Nelson-Aalen estimator, which is a non-parametric estimator of the CHF [89]. All cases within the same terminal node have the same CHF. The ensemble CHF is constructed as the average over the CHF of the B survival trees.

Noteworthy, the survival function and cumulative hazard function as linked as follows [126]:

$$S(t) = e^{-H(t)}$$

where $H(t)$ and $S(t)$ denote the cumulative hazard function and the survival function, respectively.

7.2.2 Self-training method

The semi-supervised learning (SSL) paradigm is a combination of supervised and unsupervised learning and has been widely used in many applications such as healthcare [118, 157, 6]. The primary goal of SSL methods is to take advantage of the unlabeled data in addition to the labeled data, in order to obtain a better prediction model. The acquisition of labeled data is usually expensive, time-consuming, and often difficult, specifically when it comes to healthcare and follow-up data. Hence, achieving good performance with supervised techniques is challenging, since the number of labeled instances is often too small. Over the years, many SSL techniques have been proposed [213, 195]. In this article,

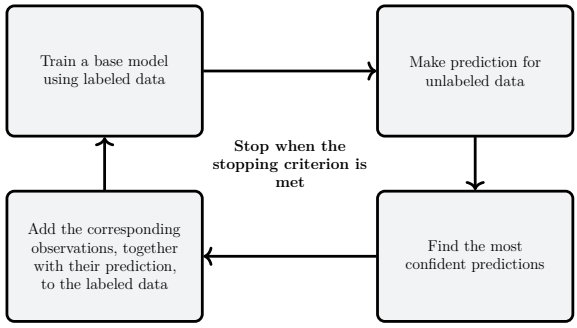


Figure 7.1: Self-training framework. The framework takes a set of labeled and unlabeled data instances as input and starts in the top left box.

we will focus on self-training (sometimes also called self-learning) [209], one of the most widely used algorithms for SSL. Self-training has been used in different approaches like deep neural networks [31], face recognition [158], and parsing [122]. This framework overcomes the issue of insufficient labeled data by augmenting the training set with unlabeled instances. It starts with training a model using a base learner on the labeled set and then augments this set with the predictions for the unlabeled instances that the model is most confident in (see Figure 7.1). This procedure is repeated until a certain stopping criterion is met. This stopping criterion, the number of instances to augment in each iteration, and the definition of confidence are instantiated according to the problem at hand.

7.3 Related work

Semi-supervised learning (SSL) methods have been applied in many different domains [213, 195]. However, few efforts have been made in order to generalize SSL algorithms to be suitable for survival analysis.

Bair and Tibshirani [5] combine supervised and unsupervised learning to predict survival times for cancer patients. They first employ a supervised approach to select a subset of genes from a gene expression dataset that correlates with survival. Then, unsupervised clustering is applied to these gene subsets to identify cancer subtypes. Once such subtypes are identified, they apply again supervised learning techniques to classify future patients into the appropriate subgroup or to predict their survival. Although the authors call the resulting approach semi-supervised, their setting is clearly different from ours.

There has also been some work that models a survival analysis task as a semi-supervised learning problem by employing a self-training strategy to predict

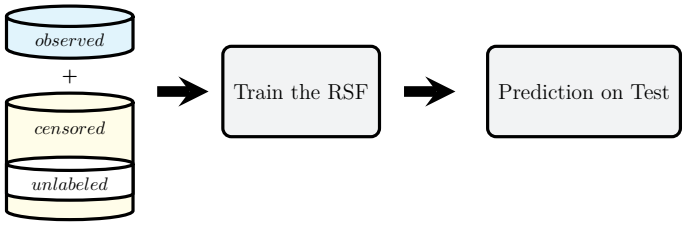


Figure 7.2: Pipeline for the first approach, called RSF+UD.

event times from observed and censored data points. Both [169, 63] treat the censored data points as unlabeled, thereby ignoring the time-to-event information that they contain. Liang et al [115] do use some information from the censored times, in the sense that they disregard data points for which the model predicts a value lower than the right-censored time points. They combine Cox proportional hazard (Cox) and accelerated failure time (AFT) model in a semi-supervised set-up to predict the treatment risk and the survival time of cancer patients. Regularization is used for gene selection, which is an essential task in cancer survival analysis. The authors found that many censored data points always violate the constraint that the predicted survival time should be higher than the censored time, restricting the full exploitation of the censored data. Therefore, in follow-up work [22], they embedded a self-paced learning mechanism in their framework to gradually introduce more complex data samples in the training process, leading to a more accurate estimation for the censored samples. An important difference between our work and the discussed studies is that we consider situations where apart from fully observed and censored instances, we also have a third category, namely extra data points that are unlabeled. To our knowledge, this is the first study to investigate the use of unlabeled instances in the survival context.

7.4 Methodology

In order to predict event times in the presence of observed, censored, and unlabeled instances, we propose three approaches.

The first approach is a straightforward application of a survival analysis method (in our case, RSF), in which we add the unlabeled set as censored instances, with the corresponding event time set to zero. We call the first approach random survival forest with unlabeled data (RSF+UD). Figure 7.2 depicts the block diagram of the first proposed pipeline.

In the second approach, we apply a semi-supervised learning approach called self-trained random survival forest (ST-RSF). In particular, we use the widely

used self-training wrapper technique [139]. Figure 7.3 shows the learning process in our self-training algorithm. This technique first builds an initial model using RSF over the labeled (in our case, observed and censored) data points and then iteratively augments the labeled set with the most confident predictions of survival time for the unlabeled dataset. In order to predict the survival time for each individual, we calculate the expected future lifetime (T_p) which at a given time t_0 is the time remaining until the event, given that the event did not occur until t_0 [126]:

$$T_p = \frac{1}{S(t_0)} \int_{t_0}^{\infty} S(t)dt \tag{7.1}$$

where $S(t)$ is the survival function predicted by RSF.

The aim is to boost the performance of the model using unlabeled data through an iterative process. As mentioned in Section 7.2, the adoption of a self-training approach requires the instantiation of three aspects. First, in order to define the confidence in a prediction, we use the variance of predictions across trees. The lower this variance, the more the trees agree, and thus, the more confidence we have in the predicted value. Second, we set the number of instances to be added to the labeled set in each iteration to 10% of the size of the unlabeled set. We set the status of these newly added instances to observed and add their predicted value as their survival time. Finally, we need to define a global stopping criterion, to terminate the iterative procedure. For this purpose, in the first iteration, we take the first quartile of the variance values and use it as the maximally allowed variance in the whole procedure. Thus, we only augment unlabeled instances if their prediction variance is smaller than this value. If no instances can be added, the algorithm stops.

The details of this approach (ST-RSF) are described in Algorithm 1.

The third approach is an adaptation of the second one, where we exploit the information contained in the censored instances to replace the arbitrarily set stopping criterion of the second approach. In particular, we use the self-training wrapper technique as before, but build the initial model over only the observed data points and iteratively augment the training set with high confidence predictions from the censored and unlabeled dataset. In other words, we treat the censored examples as unlabeled, and the observed examples as labeled, and cast the problem as a pure semi-supervised learning problem. However, in this scenario, the censored instances are not totally unlabeled, since we know that their event time is greater than the censoring time (assuming right-censored instances). As a result, we aim to exploit this information of censored instances to introduce a smarter stopping criterion in the data augmentation process.

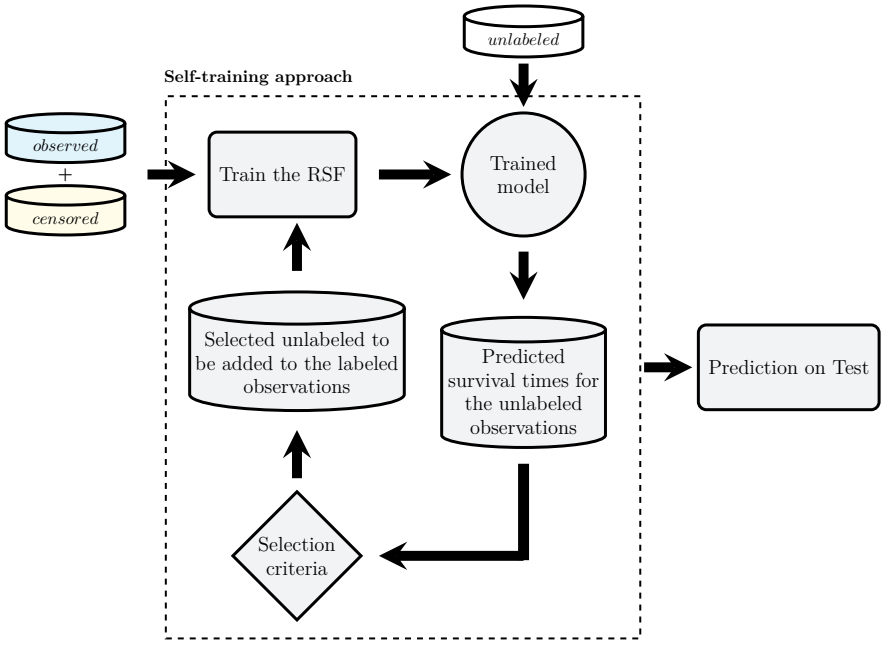


Figure 7.3: Pipeline for the second approach, called ST-RSF.

Algorithm 1: Self-trained random survival forest (ST-RSF).

Input: labeled data ($Ldata$), unlabeled data ($Udata$)
Output: Prediction model for survival time

repeat
  Train a base model using $Ldata$;
  Make a prediction for the survival time (T_p) of each instance in $Udata$ using Equation 7.1;
  Calculate the variance for each prediction;
  Sort the predictions based on minimum variance;
  If the stopping criterion is not defined yet ($S = -\infty$), find the first quartile as the stopping criterion (only in the first iteration);
  Select the top 10% $Udata$ instances from the sorted list of predictions, with variance smaller than S (confident predictions);
  Remove the confident predictions from $Udata$ and add them to the training set ($Ldata$);
until *no confident predictions have been added to the training set*;

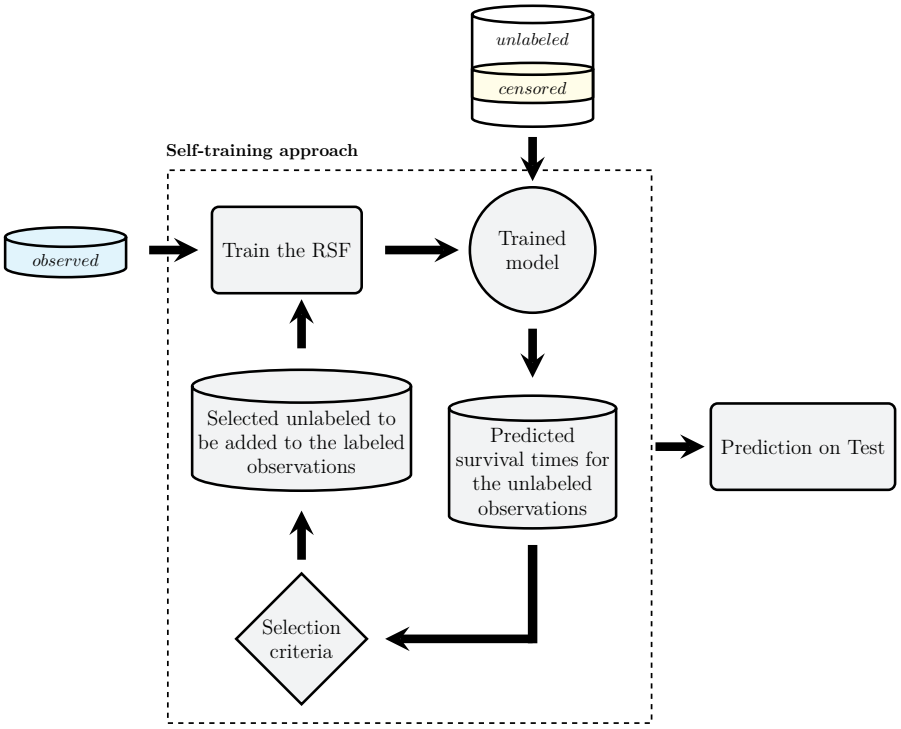


Figure 7.4: Pipeline for the third approach, called ST-RSF+CCT.

We denote this approach as a self-trained random survival forest corrected with censored times (ST-RSF+CCT). Figure 7.4 shows the learning process in this self-training algorithm.

When deciding which unlabeled (including censored) instances to add to the augmentation process, similarly to the previous approach, we assess the confidence of the ensemble predictions based on the variance of the individual tree predictions. We sort the predictions based on minimum variance (note that the resulting list contains instances both from the censored and unlabeled dataset), but instead of picking the top 10%, we use the information in the censored instances to decide when to stop adding instances. More precisely, we know that the true event time must be greater than the censoring time for those instances. As a result, whenever we encounter a censored instance with a predicted time T_p smaller than its censoring time T_c, we stop the augmentation for the current iteration. When an iteration yields zero augmented instances, the whole procedure is terminated. Preliminary experiments showed that the condition $T_c \leq T_p$ is often too strict and results in premature termination. This

Algorithm 2: Self-trained random survival forest corrected with censored times (ST-RSF+CCT).

Input: observed data (*observed*), censored data (*censored*), unlabeled data ($Udata$)
Output: Prediction model for survival time

repeat
  Train a base model using *observed*;
  Make a prediction for the survival time (T_p) of each instance in $censored \cup Udata$ using Equation 7.1;
  Calculate the variance for each prediction;
  Sort the predictions based on minimum variance;
  Calculate 95% tolerance interval corresponding to two times the standard deviation of the individual tree predictions ($T_p \pm 2\sigma$) for the instances from *censored* ;
  Find the first *censored* instance i from the sorted predictions whose censoring time (T_c) is greater than $T_p + 2\sigma$ (does not meet the criterion);
  Remove all instances sorted before i (confident predictions) from $censored \cup Udata$ and add them to the training set (*observed*);
until *no confident predictions have been added to the training set*;

happens when the prediction variances are high, and thus adding or removing some trees from the forest could result in a substantially different T_p value and thus a different condition outcome. For this reason, we calculate the 95% tolerance interval around T_p and require T_c to be smaller than or inside the tolerance interval. In other words, we allow T_c to be larger than T_p, but only if it is within its 95% tolerance interval (see Figure 7.5). For the instances (censored or unlabeled) that meet the criterion to be added to the training set, we set the status to observed with the survival time equal to T_p. Note that this removes the need to use a machine learning method that is able to work with censored instances as the base learner. In this article, in order to be consistent with and provide a fair comparison to the previous approaches, we still use RSF in this approach. The details of this approach are described in Algorithm 2.

7.5 Experimental set-up

In this section, first, we start with a dataset description and then explain the process of creating an unlabeled dataset. In Section 7.5.3, we discuss the metric of evaluation, and finally, in Section 7.5.4, we explain the comparison methods and parameter instantiation.

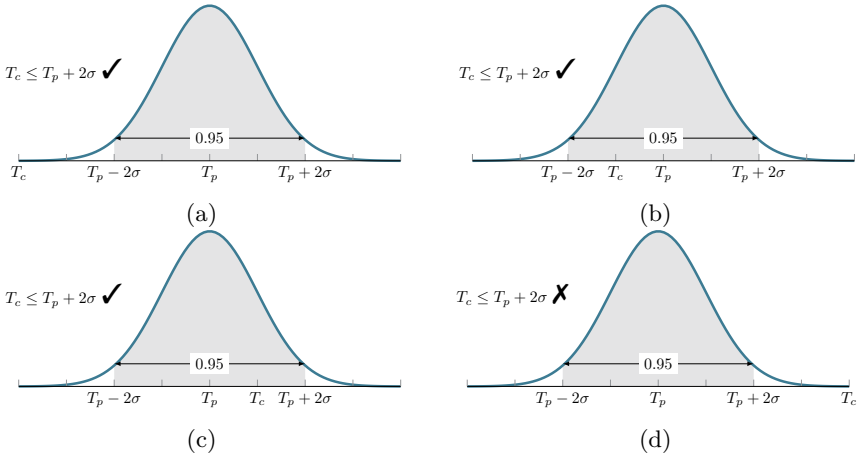


Figure 7.5: Tolerance interval corresponding to two times the standard deviation. Figures a, b, and c represent situations where the condition $T_c \leq T_p + 2\sigma$ is fulfilled, where σ is the standard deviation of the individual tree predictions, and hence, these situations are accepted by our method. In Figure d, the condition is violated.

7.5.1 Dataset description

We investigate the performance of our proposed approaches on real-life datasets from the *survival* package [185] in R as well as high-dimensional datasets from [183, 49], and some from the R/Bioconductor package. To assess the effectiveness of the proposed approaches in high-dimensional scenarios ($p \gg n$), we used ten different gene expression datasets. These datasets typically contain the expression levels of thousands of genes across a small number of samples (< 300), giving information about demographic features, disease type, survival time, etc. For convenience, in datasets with more than 10000 gene expression features, we reduced the total number of features to the top 10000 features with the largest variance across all samples. Table 7.1 shows the description and characteristics of the used datasets. The prediction task for all datasets is survival time (time to death).

7.5.2 Unlabeled data generation

Since we are not aware of survival datasets that include unlabeled instances, we artificially remove the label of a subset of instances as follows (see Figure 7.6). First, we take the original data and construct five folds for cross-validation, in order to have a fair evaluation of our approach. Then, for each training set in the cross-validation (i.e., for each combination of four folds), we construct the unlabeled category. We split the training data into two sets called labeled data

Table 7.1: Characteristics of the used clinical and high-dimensional datasets.

Name	#Observations	#Features	Censoring rate
Veteran	137	6	6%
Lung	228	8	27%
PBC	312	17	60%
DrAsGiven	119	22122	42%
EMTAB386	129	10364	44%
GSE14764	80	13112	74%
GSE32062	260	20112	54%
Norway/Stanford Breast Cancer Data (NSBCD)	115	549	67%
Sporadic lymph-node-negative patients (Veer)	78	4751	56%
Dutch Breast Cancer Data (DBCD)	295	4919	73%
Diffuse Large-B-Cell Lymphoma data (DLBCL)	240	7399	42%
Lung adenocarcinomas (LungBeer)	86	7129	72%
Acute myeloid leukemia (AML)	79	54675	40%
Breast invasive carcinoma (BRCA)	1080	117	86%
First National Health and Nutrition Examination Survey (NHANES I)	9549	21	64%

(Ldata) and unlabeled data (Udata). To have a fair and accurate evaluation, we make sure to have the same distribution for both sets relative to the status (being censored or observed). Then, we take Udata and make the instances unlabeled by removing their time and status values.

To improve the stability of the results, we repeat the cross-validation process 10 times and report the average results. We also vary the percentage of unlabeled instances from 5% to 75% of the original training set.

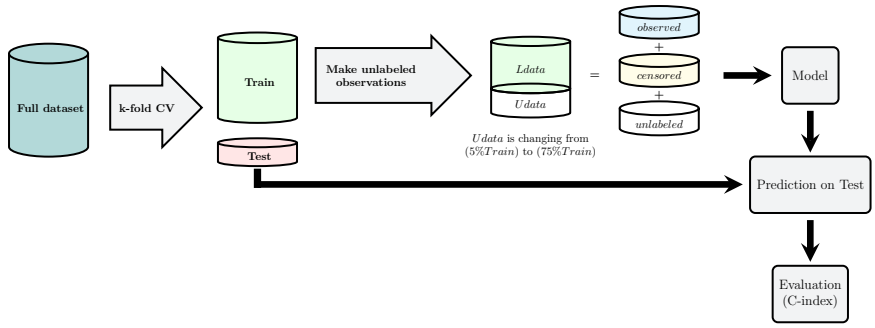


Figure 7.6: Illustration of the used procedure in the chapter. The first part illustrates the process of making an unlabeled set. Then, the box Model uses one of the three proposed approaches. Predictions are made for the Test set, and finally, evaluations are made using the evaluation metric (C-index).

7.5.3 Performance evaluation

In survival analysis, instead of measuring the absolute survival time for each instance, a popular way to assess a model is to estimate the relative risk of an event occurring for different instances. The Harrell's concordance index (C-index) [61] is a common way to evaluate a model in survival analysis [165]. C-index can be interpreted as the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered. In other words, it is the probability of concordance between the predicted and the observed survival time. Two subjects' survival times can be ordered not only if (1) both of them are observed but also if (2) the observed time of one is smaller than the censored survival time of the other [180]. Consider a set of observation and prediction values for two different instances, $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$, where y_i and $\hat{y}_i$ represent the actual survival time and the predicted value, respectively. The concordance probability between these two instances can be computed as $c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 > y_2)$. In this chapter, we compute the C-index for each test fold in the cross-validation process and return the average value over the 5 test folds.

7.5.4 Comparison methods and parameter instantiation

We applied five different methods: the three methods presented in this article, namely RSF+UD, ST-RSF and ST-RSF+CCT, and standard RSF and Lasso-Cox trained on the Ldata set only. The goal to perform RSF was to address the first research question (see Section 7.1), i.e., to investigate if adding an unlabeled set to the training phase would increase the performance of the model. The comparison of the three proposed approaches addresses the second research question. To avoid falling into a slightly biased random survival forest comparison, we have reported results of Cox regression with LASSO regularization (Lasso-Cox) as a baseline model. Lasso-Cox introduces the $L1$ norm penalty in the Cox log-likelihood function [188]. Since the majority of our used datasets are high-dimensional ($p \gg n$), we have employed Lasso-Cox due to its capability of handling high-dimensional datasets.

In order to estimate the generalization capacity of the models, we performed a 5-fold cross-validation on each dataset and estimate test accuracy, and repeated it 10 times to achieve reliable results. It is worthwhile to mention that the optimal tuning parameter (λ) in Lasso-Cox is chosen by nested cross-validation while no hyperparameter tuning has been employed for the other approaches. For RSF-based methods, the number of trees was set to 500, and the number of candidate variables considered in each tree node was set to $p/3$, where p is the number of variables.

7.6 Results and discussion

Figures 7.7 and 7.8 show the performance of the methods, for different percentages of labeled instances for twelve datasets from the fifteen. For each figure, we show six different curves. The blue and dark green curves represent the performance of RSF and Lasso-Cox using only labeled data (Ldata), respectively. The orange line (maximum) shows the performance of RSF using the complete training set as labeled data and is included as a reference to see how much performance we could gain by having access to all (observed or censored) information. The other three curves represent the proposed approaches.

The figures show that the performance of RSF can indeed be improved by adding unlabeled data to the training set. There are often big performance gains, especially with a lower percentage of labeled instances; however, this improvement does not hold for all datasets and all approaches.

From the figures, we can see that ST-RSF+CCT is the best approach overall, although it often starts in the second or even third position with very few labeled examples. This could be due to a lack of sufficient censored data to guide the augmentation process. Although in three datasets, it starts at a performance lower than RSF, on datasets with a very small number of samples (e.g., Veer, LungBeer, and GSE14764 all with less than 100 observations), ST-RSF+CCT is immediately much better than RSF. Note that 25% of labeled instances can be as low as 15 labeled examples for the Veer dataset, where ST-RSF+CCT outperforms RSF in the first part of the graph, reaching a C-index level of around 69%.

In addition, in several datasets with a high percentage of censored instances (e.g., DBCD, GSE14764, EMTAB, NHANES I, BRCA, and Veer, all with higher than 43% censoring rate), ST-RSF+CCT is performing as the best method in almost all percentages of labeled instances.

When comparing the curves for ST-RSF and ST-RSF+CCT, we see in the majority of datasets that either ST-RSF+CCT is on the winning hand over the entire curve, or ST-RSF is better in only some parts. In addition, in most cases, C-index values for ST-RSF fluctuate when changing the number of labeled instances; however, ST-RSF+CCT shows more steady behavior by feeding more labeled instances. Moreover, when comparing the range of C-indices (difference between min and max), ST-RSF varies more dramatically in most experiments; but overall, ST-RSF+CCT acts robustly. This could be due to the fact that for censored instances, ST-RSF+CCT compares the predicted survival time with the censoring time, which results in more confident predictions.

While one would expect the largest gain from using unlabeled data in settings where very few labeled data are available, we see that also considerable

improvements can be obtained at the other extreme, where most training instances are labeled and only a small portion, say 5 or 10%, of unlabeled instances are added. Especially the self-training approaches seem to achieve good results compared to RSF there, although the variability is high. This raises the question of how these techniques would compare to RSF in regular survival analysis tasks (i.e., without an unlabeled set) and can be an interesting direction for future work.

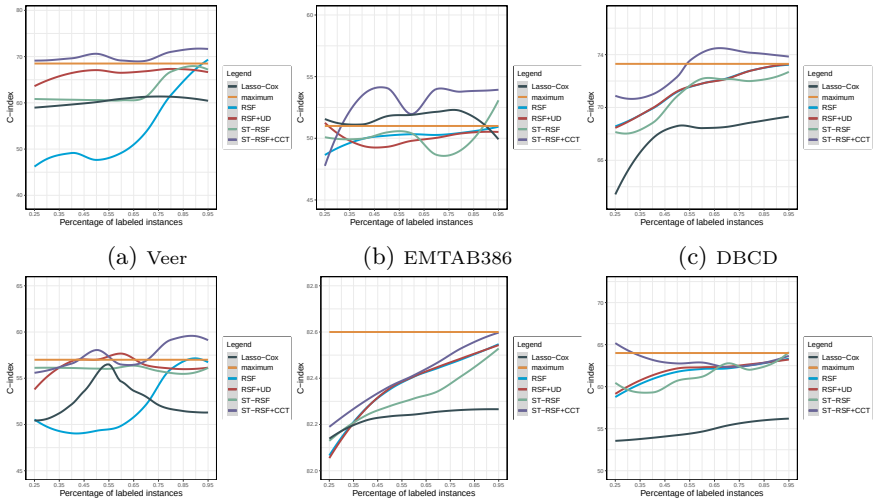
A related observation is that the proposed approaches (especially the semi-supervised ones) are able to beat the 'maximum' performance on several occasions. This demonstrates that they are able to select the most reliable instances and leave instances that can harm predictive performance (e.g., noisy instances) out of the training set.

When looking at the RSF+UD curve, we see that it often closely follows the RSF curve for a substantial part of the graph (e.g. for the datasets PBC, GSE32062, DBCD, NHANES I, and BRCA). This is due to the fact that the resulting ensembles are very similar. In fact, the trees generated by RSF are contained in the trees generated by RSF+UD, since the addition of censored data points with event time set to zero does not influence the log-rank splitting criterion, but only alters the size of the trees.

Since the visual inspection of the figures makes it difficult to draw strong conclusions, we also conducted a more aggregated comparison by comparing the areas under the plotted curves. Table 7.2 shows the means and standard deviations of the AUC rate on the datasets, as well as the average accuracy of each algorithm. As can be seen in Table 7.2, all our proposed methods provide better results than RSF for all datasets. More specifically, the third approach (ST-RSF+CCT) outperforms RSF, ST-RSF, and Lasso-Cox and manages to be statistically significantly better according to the Friedman-Nemenyi test (Figure 7.9) [41][1]. The second best method, on average, is the RSF+UD variant, which also statistically significantly outperforms RSF and Lasso-Cox, and has a slight, non-significant, margin over ST-RSF. Furthermore, based on the reported results in Table 7.2, in all high-dimensional datasets, either ST-RSF+CCT or ST-RSF are the winning algorithms, meaning that both proposed algorithms are performing better in high-dimensional settings.

The use of self-training approaches may raise concerns related to overfitting. Since our algorithms have no tunable hyperparameters, they are not prone to the kind of overfitting that results from the hyperparameter tuning process in other algorithms. Moreover, random forests overall are known to be robust

[1] In a critical distance diagram, those algorithms that are not joined by a line (i.e., their rankings differ more than a critical distance (CD)) can be regarded as statistically significantly different [41].



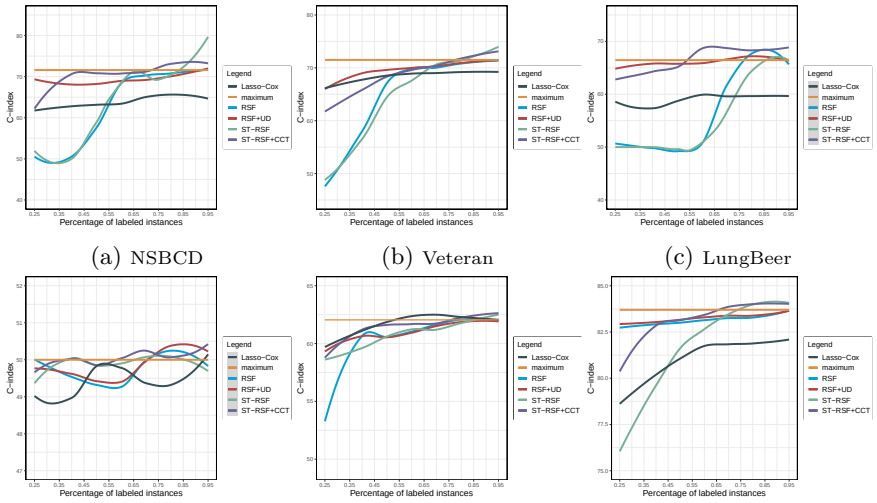
(a) Veer (b) EMTAB386 (c) DBCD

(d) GSE14764 (e) NHANES I (f) BRCA

Figure 7.7: Evaluation of the performance of the methods, for different percentages of labeled instances for six datasets with a high percentage of censored instances.

to overfitting [17], due to the fact that by increasing the number of trees, the variance of the error gets reduced. Nevertheless, we have investigated the learning curve of the ST-RSF and ST-RSF+CCT algorithms on two datasets with 55% labeled instances (see Figure 7.10). We have chosen NSBCD and Veteran, as they have different censoring rates (67% versus 6% censoring rate). In Figure 7.10, the numbers indicated on the training curves show the number of augmented instances in each step. Due to the difference in the augmentation process, in each iteration, ST-RSF+CCT augments fewer instances compared to ST-RSF. For instance, for Veteran, from 42 unlabeled instances, after six iterations (when the stopping conditions hold), ST-RSF has augmented 29 instances, in comparison to 11 for ST-RSF+CCT. The difference in the number of augmentations for ST-RSF and ST-RSF+CCT confirms that the third approach is more conservative and hence leads to less overfitting.

Our findings demonstrate, first, that adding unlabeled data to the training set enhances the performance of the algorithm (cfr. our first research question), and second, that from the approaches that we have proposed, the self-training technique that uses the information in the censored data points to guide the data augmentation process performs best (cfr. our second research question). The concept of the idea that we proposed could be applied using other base learners and semi-supervised learning strategies, but it remains to be investigated whether the results carry over to other learners.



(a) NSBCD (b) Veteran (c) LungBeer

(d) GSE32062 (e) Lung (f) PBC

Figure 7.8: Evaluation of the performance of the methods, for different percentages of labeled instances for six datasets.

7.7 Conclusion

In this article, we have investigated the inclusion of unlabeled data points in a survival analysis task. More precisely, we have considered learning from data with three degrees of supervision: fully observed, partially observed (censored), and unobserved (unlabeled) data points. To our knowledge, this is a setting that has not been considered before. We have proposed three different approaches for this task. The first approach treats the unlabeled points as censored and applies a standard survival analysis technique. The second one applies a standard semi-supervised wrapper approach on top of a survival analysis task. The third one is an adaptation of the second, which treats the censored instances as unlabeled

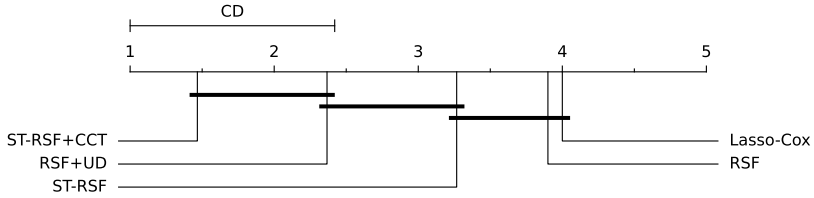
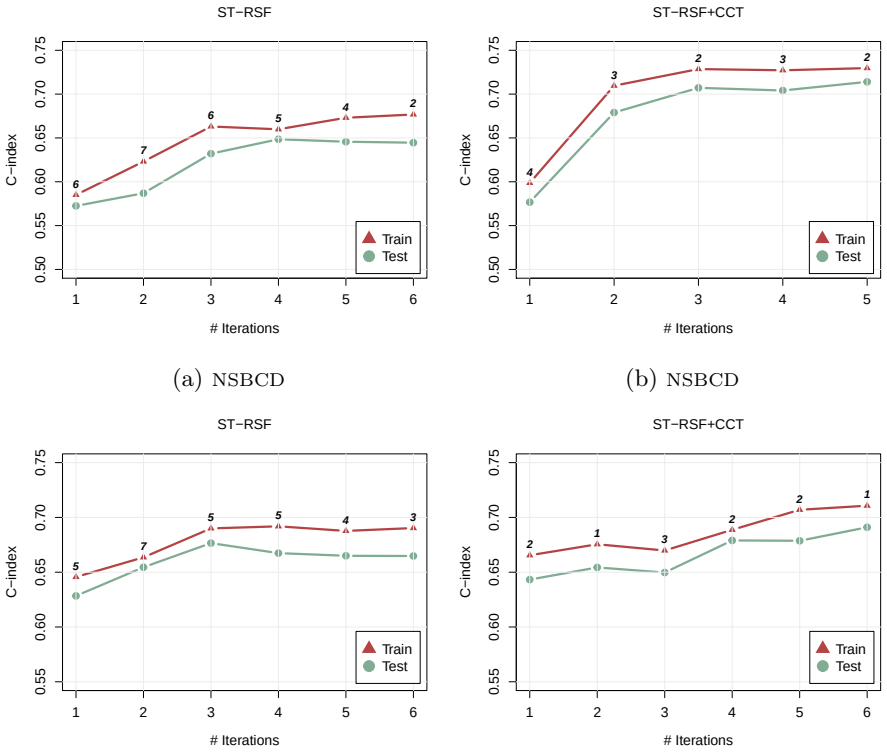


Figure 7.9: Results of the Friedman-Nemenyi test of methods ranking. The five methods are compared in terms of their ranking using the evaluation measure, AUC.



(a) NSBCD (b) NSBCD

(c) Veteran (d) Veteran

Figure 7.10: Learning curves of the ST-RSF+CCT and ST-RSF methods for NSBCD (figures a and b) and Veteran (figures c and d) datasets. The plots have been shown for 55% labeled data for both datasets.

but manages to exploit the censored information to guide the semi-supervised approach. We have evaluated and compared the proposed approaches on fifteen real-world survival analysis datasets, including clinical and high-dimensional ones. Our results have shown that, first, adding unlabeled instances to the training set improves the predictive performance on an independent test set. Second, the third proposed approach generally outperforms the others due to its ability to integrate partial supervision information inside a semi-supervised learning approach.

Our findings can be quite helpful, especially in the healthcare area, where studies often require long-term follow-up of patients, which is costly and challenging. For instance, for the prediction of long-term outcomes after hospitalization,

Table 7.2: Performance in terms of Area Under the Curve (AUC).

Datasets	Lasso-Cox	RSF	RSF+UD	ST-RSF	ST-RSF+CCT
Veteran	68.4 ± 5.1	64.71 ± 5.1	$\mathbf{69.7} \pm 5.7$	64.33 ± 6.07	68.71 ± 5.9
Lung	$\mathbf{61.7} \pm 4.64$	60.55 ± 5.3	61.05 ± 5.3	60.79 ± 5.5	61.55 ± 8.9
PBC	81.05 ± 3.8	83.12 ± 3.4	$\mathbf{83.24} \pm 3.4$	81.67 ± 3.6	83.19 ± 3.4
BRCA	54.81 ± 5.6	61.68 ± 5.4	61.93 ± 5.3	61.1 ± 6.5	$\mathbf{63.14} \pm 5.8$
NHANES I	82.23 ± 0.62	82.37 ± 0.62	82.37 ± 0.62	82.32 ± 0.64	$\mathbf{82.41} \pm 0.7$
DrAsGiven	52.27 ± 7.9	53.46 ± 7.8	55.3 ± 8.7	$\mathbf{56.18} \pm 5.2$	53.13 ± 1.8
EMTAB386	51.54 ± 7.6	50.14 ± 7.2	50.05 ± 8.02	50.12 ± 4.02	$\mathbf{52.83} \pm 7.2$
GSE14764	51.14 ± 14.3	52.14 ± 13.7	56.38 ± 16.2	56.65 ± 16.1	$\mathbf{57.52} \pm 16.1$
GSE32062	49.37 ± 5.4	49.76 ± 6.3	49.87 ± 6.4	49.93 ± 6.1	$\mathbf{50.04} \pm 6.2$
NSBCD	63.88 ± 8.9	62.41 ± 7.4	69.34 ± 8.1	64.65 ± 7.2	$\mathbf{70.85} \pm 7.6$
Veer	60.41 ± 10.5	54.45 ± 15.6	66.49 ± 10.1	62.98 ± 10.6	$\mathbf{70.09} \pm 9.8$
DBCD	67.94 ± 5.6	71.37 ± 5.3	71.38 ± 5.3	70.77 ± 5.5	$\mathbf{73.10} \pm 5.7$
DLBCL	58.12 ± 5.2	59.82 ± 5.3	58.94 ± 5.3	$\mathbf{60.26} \pm 5.4$	59.82 ± 5.3
LungBeer	58.94 ± 13.2	56.93 ± 7.8	66.16 ± 12.7	57.21 ± 4.8	$\mathbf{66.67} \pm 10.5$
AML	59.06 ± 8.6	53.3 ± 8.7	59.47 ± 9.9	55.80 ± 7.8	$\mathbf{60.32} \pm 9.4$
Average	61.39	61.01	64.17	62.32	**64.89**

our results suggest that the study data could be complemented by additional routinely collected baseline data available in the hospital database management system, from patients matching the inclusion and exclusion criteria, but not included in the follow-up study. Moreover, based on the results, our proposed algorithms (ST-RSF+CCT and ST-RSF) perform better in high-dimensional settings (gene-expression datasets) which is a common dataset type in the healthcare area.

A limitation of our study is that our experiments assume that the unlabeled set is a random subset of the labeled dataset where the labels have been removed, leading to no trend or bias in the unlabeled set. When employed in a clinical setting, the unlabeled set should be carefully provided to not incorporate a biased set so that the procedure does not introduce noise through these additive iterations in the algorithm.

Declarations

Funding. No funding was received to assist with the preparation of this manuscript

Conflicts of interest/Competing interests. The authors have no conflicts of interest to declare that are relevant to the content of this article

Ethics approval. Not applicable

Consent to participate. No tests, measurements, or experiments were performed on humans as part of this work

Consent for publication. The authors have agreed to submit it in its current form for consideration for publication in Journal

Availability of data and material. All of the datasets used in this article are publicly available and have been referenced

Code availability. The source code will be publicly available

Authors' contributions.

Fateme Nateghi Haredasht: Investigation, Methodology, Software, Writing, Original draft.

Celine Vens: Supervision, Conceptualization, Methodology, Writing, Review and editing

Acknowledgements

This work was supported by KU Leuven Internal Funds (grant 3M180314). The authors also acknowledge the Flemish Government (AI Research Program).

Chapter 8

Exploiting censored information in self-training for time-to-event prediction

The following chapter is submitted to the International Journal of Medical Informatics:

Nateghi Haredasht, F., Dauda, K.A. & Vens, C. (2022). Exploiting censored information in self-training for time-to-event prediction.

Abstract

A common problem in medical applications is predicting the time until an event of interest. Traditionally, classical survival analysis techniques have been used to address this problem. However, these techniques are of limited usage when considering nonlinear and interaction effects among biomarkers, and high profiling survival datasets. Although supervised machine learning techniques have shown some advantages over standard statistical methods in handling high-dimensional datasets, their application to survival analysis is at best limited. A major reason behind this is the difficulty in processing censored data, which is a common component of survival analysis. In this paper, we have transformed the time-to-event prediction problem into a semi-supervised regression problem in which we use a self-training wrapper approach with random survival forests as the base learner. In this approach, censored observations are introduced as partially labeled observations since their predicted time (target value) should exceed the censoring time. First, the algorithm builds a base model over the observed instances and then augments them iteratively with highly confident predictions over the censored set, using a smart stopping criterion based on the censoring time. The proposed approach has been evaluated and compared on fifteen real-world survival analysis datasets, including clinical and high-dimensional data. The ability of our proposed approach to integrate partial supervision information within a semi-supervised learning strategy has enabled it to achieve competitive performance compared to baseline models, particularly in the case of a high-dimensional regime.

8.1 Introduction

Survival analysis is a subfield of statistics concerned with the analysis of data where the outcome of interest is the time until a particular event of interest occurs. There is a widespread use of survival analysis in medicine, where events of interest might include death, tumor recurrence, and hospital discharge, among others. Censoring, which can occur for various reasons such as drop-out, is one of the main challenges of survival analysis. Observations that are censored (right-censored or left-censored) cannot provide the true survival time, as, for example, in the right-censored case, we know that the observed time is an underestimate of the survival time [35].

Traditionally, methods like Cox Proportional Hazards (CPH) and Accelerated Failure Time (AFT) models have been widely used throughout literature to overcome censoring; however, these methods have been unable to cope with real-world datasets with hundreds or thousands of features. Additionally, these models are not able to incorporate the nonlinear relationship that exists between the features [216, 38].

The field of survival analysis has adopted many supervised machine learning algorithms in recent years, but the problem of applying these techniques directly to censored data is challenging, since the time to event is only partially known. Sometimes the task is transformed into a binary classification task (does the event happen before a certain time?), in which censored data points are either eliminated [196] or their impact is diminished by a weighting procedure [215]. A number of machine learning algorithms have been successfully modified to employ censored information in survival analysis. For example, decision trees [56], artificial neural networks (ANN) [46], and support vector machines (SVM) [96]. Among the most popular ensemble-based frameworks are bagging survival trees [77] and random survival forests [84]. There has also been an extension of more advanced learning tasks such as active learning [198] and transfer learning [114] towards survival analysis.

Although in recent years, applying supervised machine learning-based techniques in the survival analysis domain has gained attention [202], semi-supervised learning (SSL) methods [213, 195] are also briefly addressed in the survival analysis literature. The study by Bair and Tibshirani [5], combines supervised and unsupervised learning to predict survival times for cancer patients. A supervised approach is used to select a subset of genes from a gene expression dataset that correlates with survival. Then, to identify cancer subtypes, unsupervised clustering is applied to these gene subsets. Having identified such subtypes, they apply supervised learning techniques again to classify future patients into the appropriate subgroups (low-risk or high-risk) or to predict their survival. Using the median survival time, they created two classes. The

"low-risk" group of patients would be those who lived longer than the median survival time, while the "high-risk" group would be those who died sooner than the median survival time. For the censored patients, based on the Kaplan-Meier survival curve for all the patients, they estimate the probability that a censored case survives a specified length of time and thus belongs to the "low-risk" and "high-risk" classes, respectively.

Furthermore, there has been some research that models a survival analysis task as a semi-supervised learning problem by employing a self-training strategy to predict event times from observed and censored data. Both [169, 63] treat the censored data points as unlabeled, thus not taking into account the time-to-event information that these data points provide.

Liang et al [115] do use some information from the censored times, in the sense that they disregard data points for which the model predicts a value lower than the right-censored time points. They combine Cox proportional hazard (Cox) and accelerated failure time (AFT) model in a semi-supervised set-up to predict the treatment risk and the survival time of cancer patients. Regularization is used for gene selection, which is an essential task in cancer survival analysis. The authors found that many censored data points consistently violate the constraint that the predicted survival time should be higher than the censored time, restricting the full exploitation of the censored data. Therefore, in follow-up work [22], they embedded a self-paced learning mechanism called Cox-SP-AFT in their framework to gradually introduce more complex data samples in the training process, leading to a more accurate estimation for the censored samples. To estimate the coefficients of the AFT model, they introduce a loss function derived from the constraint that the survival time must not be less than the censoring time. As a result, if the estimated survival time of a sample is less than the censoring time, then this sample must be falsely labeled, and its loss value must be positive infinity. A censored sample, however, has a square loss function if it obeys the censoring condition. Then in order to select confident samples from the censored dataset, they define a threshold (age parameter) for the loss function in which the samples with losses smaller than the age parameter (α) will be kept at the training phase, otherwise will be assigned zero weight. This technique was intended to be used to classify cancer patients.

In a separate study, Roy et al [160] modeled the time-to-event prediction as a multi-target regression problem, with censored observations modeled as partially labeled. More specifically, the different event times in the dataset are viewed as binary targets. For each data instance, it is specified whether it has experienced the event or not at each time stamp, using missing values when an instance has been censored after a certain period of time. Then they apply semi-supervised predictive clustering trees and ensembles thereof to the resulting data.

Several previous studies have applied semi-supervised learning approaches to survival data analysis; however, none have utilized the underlying information contained within the censored data, which is the fact that the target value for right-censored observations should be greater than the censoring time. In this paper, using a semi-supervised learning approach, we propose a new time-to-event prediction algorithm. Specifically, this paper utilizes the widely used self-training wrapper technique [209, 113], which builds a classifier/regressor over the labeled (in our case, observed) data points and then augments the labeled set iteratively with highly confident predictions over the unlabeled (censored) set of data. Our approach uses random survival forests as the base learner [84] and compares the proposed algorithm's predictive performance with three competing methods based on fifteen real-life healthcare datasets.

The outline of the chapter is as follows. Section 8.2 introduces the background and reviews some concepts of the employed models including random survival forest and self-training approaches. Section 8.3 introduces our proposed method. Section 8.4 presents the experimental set-up, including dataset description, performance evaluation, and comparison methods and parameter instantiation. Results are presented in Section 8.5. Conclusions are drawn in Section 8.6.

8.2 Background

Before explaining the proposed method, we first start by reviewing the models that have been employed in our approach. This section discusses the random survival forest model, followed by a discussion of self-training models.

8.2.1 Random survival forest model

Random survival forests are well-known ensemble-based learning models that have been widely applied to many survival analysis applications and have been shown to outperform traditional survival analysis methods [125]. The random survival forest (RSF) [84] is quite similar to Breiman's original random forest [17]. Based on tree-structured models, the random forest algorithm makes a prediction. In a similar manner to random forests, RSF combines bootstrapping, tree building, and prediction aggregation. However, in the splitting criterion to grow a tree and in the predictions returned in the leaf nodes, RSF explicitly considers survival time and censoring information. RSF has three main steps. As the first step, it draws B bootstrap samples from the original data. In the second step, for each bootstrap sample, a survival tree is grown. At each node of a tree, p candidate variables are randomly selected, where p is a parameter, often defined as a proportion of the original number of variables. The task is to split the node into two child nodes using the best candidate variable and split point, as determined by the log-rank test [166]. The best split is the one

that maximizes survival differences between the two child nodes. Growing the obtained tree structure is continued until a stop criterion holds (e.g., until the number of observed instances in the terminal nodes drops below a specified value). In the last step, the cumulative hazard function (CHF) associated with each terminal node in a tree is calculated by the Nelson-Aalen estimator, which is a non-parametric estimator of the CHF [89]. All cases within the same terminal node have the same CHF. The ensemble CHF is constructed as the average over the CHF of the B survival trees.

Noteworthy, the survival function and cumulative hazard function are linked as follows [126]:

$$S(t) = e^{-H(t)}$$

where $H(t)$ and $S(t)$ denote the cumulative hazard function and the survival function, respectively.

8.2.2 Self-training model

As a combination of supervised and unsupervised learning, semi-supervised learning (SSL) has been used in many applications [118, 157, 6]. In order to obtain a more accurate prediction model, SSL methods seek to make use of unlabeled data as well as labeled data. In some applications, it is difficult to achieve good performance with supervised techniques due to the relatively small number of labeled instances. This is due to the fact that labeling techniques are generally expensive and time-consuming. Consequently, over the years, many SSL techniques have been proposed [213, 195]. In this article, we will focus on self-training (also called self-learning) [209], one of the most widely used algorithms for SSL. Self-training has been used in different ways like deep neural networks [31], face recognition [158], and parsing [122]. By augmenting the training set with unlabeled instances, this framework overcomes the problem of insufficient labeled data. The process begins by training a model using a base learner on the labeled data set, after which it augments the labeled data set with predictions for unlabeled instances that the model is most confident in (see Figure 8.1). The process is repeated until a certain stopping criterion is achieved. Depending on the problem at hand, the stopping criterion, the number of instances to be augmented in each iteration, and the definition of confidence are determined. When using the self-training method, the choice of each of these three steps plays an important role, especially since the first two steps are often set arbitrarily or with costly parameter optimizations.

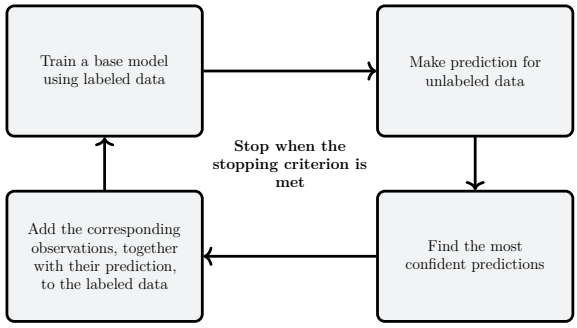


Figure 8.1: Self-training framework. The framework takes a set of labeled and unlabeled data instances as input and starts in the top left box.

8.3 The proposed method

In our proposed approach, we apply a semi-supervised learning approach, the widely used self-training wrapper technique, that was explained earlier [139].

Using the self-training wrapper technique, we build the initial model using only the observed data points, then iteratively augment it with high confidence predictions from the censored data. In other words, we treat the censored examples as unlabeled, and the observed examples as labeled, and cast the problem as a pure semi-supervised learning problem. However, in this scenario, the censored instances are not totally unlabeled, since we know that their event time is greater than the censoring time (assuming right-censored instances). As a result, we aim to exploit this information of censored instances to introduce a smarter stopping criterion in the data augmentation process. We denote this approach as STUART: Self-Trained sUrvivAl foResT which is a self-trained random survival forest corrected with censored times. Figure 8.2 shows the learning process in this self-training algorithm. This technique first builds an initial model using RSF over the labeled (in our case, observed) data points and then iteratively augments the labeled set with the most confident predictions of survival time for the unlabeled dataset (censored). In order to predict the survival time for each individual, we calculate the expected future lifetime (T_p) which at a given time t_0 is the time remaining until the event, given that the event did not occur until t_0 [126]:

$$T_p = \frac{1}{S(t_0)} \int_{t_0}^{\infty} S(t)dt \tag{8.1}$$

where $S(t)$ is the survival function predicted by RSF.

Using the variance of the individual tree predictions as a confidence measure of the ensemble predictions, we determine which unlabeled (censored) instances to

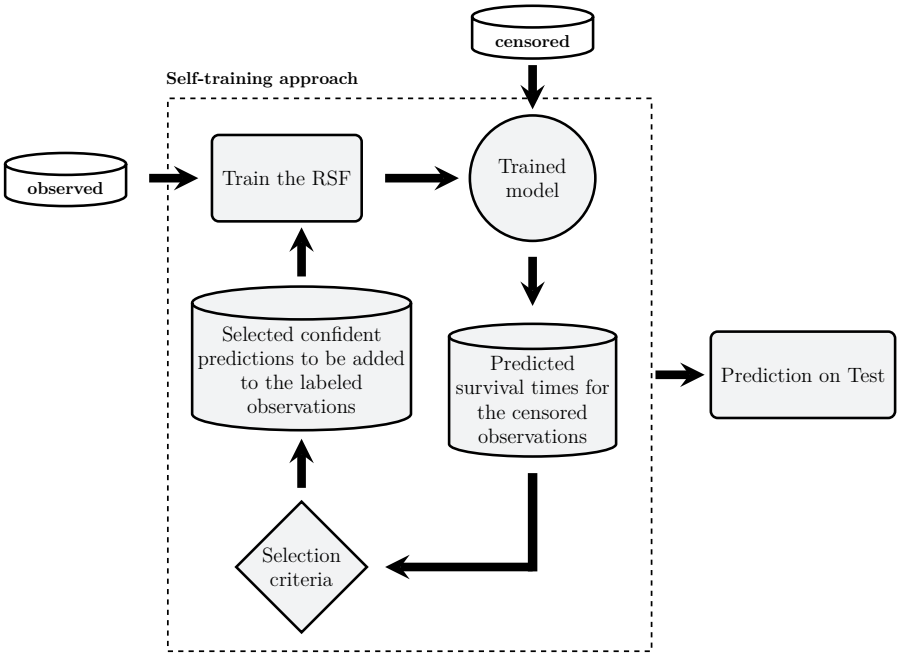


Figure 8.2: Pipeline for the proposed approach, called STUART.

add to the augmentation process. We sort the predictions in increasing order according to the variance, and then we decide when to stop adding any new instances based on the information in the censoring time. In more detail, we know that the true event time must exceed the censoring time. We, therefore, stop the augmentation process whenever a censored instance is encountered with a predicted time T_p that is lower than its censoring time T_c. If at the end of an iteration, no instances can be augmented, the entire process is terminated. In order to avoid premature termination (prediction variances can be high, in which case adding or removing some trees from the forest could result in a substantially different T_p value and therefore, a different condition outcome), we relax the condition $T_c \leq T_p$ as follows. We calculate a 95% tolerance interval around T_p and require T_c to be smaller than or within the tolerance interval. In other words, we allow T_c to be larger than T_p, but only if it is within its 95% tolerance interval (see Figure 8.3). In the case of censored examples that meet the criteria for being added to the training set, we set their status to observed with a survival time equal to T_p. Algorithm 3 describes this approach in detail. We chose RSF as the base learner, as they provide a natural way to compute the prediction confidence. Moreover, they benefit from advantages inherent in random forest techniques: high accuracy, efficient learning times, parallelizable,

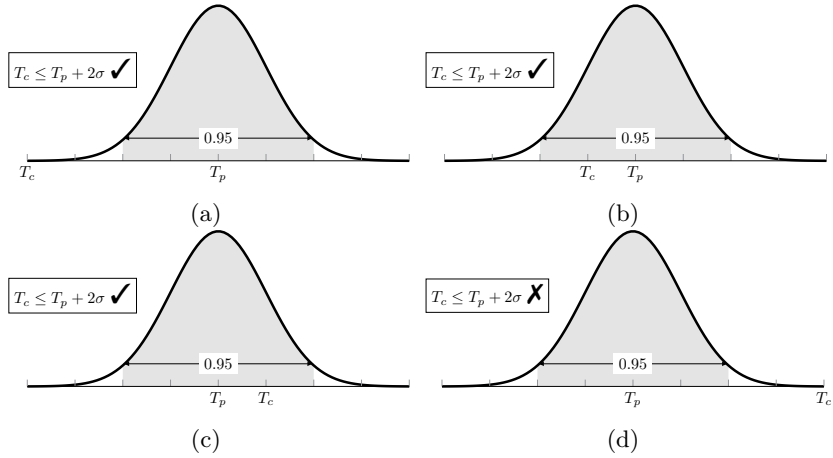


Figure 8.3: Tolerance interval corresponding to two times the standard deviation. Figures a, b, and c represent situations where the condition $T_c \leq T_p + 2\sigma$ is fulfilled, where σ is the standard deviation of the individual tree predictions, and hence, these situations are accepted by our method. In Figure d, the condition is violated.

Algorithm 3: STUART.

Input: observed data (*observed*), censored data (*censored*)
Output: Prediction model for survival time

repeat
  Train a base model using *observed*;
  Make a prediction for the survival time (T_p) of each instance in *censored* ;
  Calculate the variance for each prediction;
  Sort the predictions based on minimum variance;
  Calculate 95% tolerance interval for the predictions ;
  Find the first instance i from the sorted predictions whose censoring time (T_c) is greater than $T_p + 2\sigma$ (does not meet the criterion);
  Remove all instances sorted before i (confident predictions) from *censored* and add them to the training set (*observed*);
until *no confident predictions have been added to the training set*;

feature importance scores, etc. In addition, in the result section, we compare our method to RSF which is currently one of the state of the art methods in the survival analysis domain, and as a result, a different base learner complicates the interpretation of the results.

Table 8.1: Characteristics of the used clinical and high-dimensional datasets.

Name	#Observations	#Features	Censoring rate
Veteran [185]	137	6	6%
Lung [185]	228	8	27%
PBC [185]	312	17	60%
DrAsGiven [183]	119	22122	42%
EMTAB386 [183]	129	10364	44%
GSE14764 [183]	80	13112	74%
GSE32062 [183]	260	20112	54%
Norway/Stanford Breast Cancer Data (NSBCD) [183]	115	549	67%
Sporadic lymph-node-negative patients (Veer) [183]	78	4751	56%
Dutch Breast Cancer Data (DBCD)[183]	295	4919	73%
Diffuse Large-B-Cell Lymphoma data (DLBCL) [183]	240	7399	42%
Lung adenocarcinomas (LungBeer) [183]	86	7129	72%
Acute myeloid leukemia (AML) [183]	79	54675	40%
Breast invasive carcinoma (BRCA) [146]	1080	117	86%
First National Health and Nutrition Examination Survey (NHANES I) [49]	9549	21	64%

8.4 Experimental set-up

In this section, we first describe the datasets in detail in Section 8.4.1, then we discuss the evaluation metrics in Section 8.4.2, and we continue with an explanation of the comparison methods and parameter instantiation in Section 8.4.3.

8.4.1 Dataset description

We investigate the performance of our proposed approach on real-life datasets with different characteristics from the *survival* package [185] in R as well as high-dimensional datasets with large numbers of observations from[49], and some from the R/Bioconductor package. We also used ten different high-dimensional gene expression datasets ($p \gg n$) [183]. In these datasets, thousands of genes are typically expressed across a few samples (< 300), contributing information about demographic characteristics, disease type, survival time, etc. As it was computationally expensive to run all the competitor methods on datasets with more than 10,000 gene expression features, we reduced the number of features to the top ten thousand features with the largest variance across all samples. Table 8.1 provides a description and characteristics of the datasets used in this study. The predicted outcome for all datasets is survival time (time until death).

8.4.2 Performance evaluation

Survival Analysis involves estimating the relative risk of an event occurring for different instances rather than measuring the absolute survival time for each instance. The Harrell's concordance index (C-index) [61] is a common way to evaluate a model in survival analysis [165]. As an interpretation of the C-index, it can be defined as the fraction of all pairs of subjects whose predicted survival times are ordered correctly among all subjects whose survival times can be predicted. In other words, it is the probability of concordance between the predicted and the observed survival time. Two subjects' survival times can be ordered not only if (1) both of them are observed but also if (2) the observed time of one is smaller than the censored survival time of the other [180]. Consider a set of observation and prediction values for two different instances, $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$, where y_i and $\hat{y}_i$ represent the actual survival time and the predicted value, respectively. The concordance probability between these two instances can be computed as $c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 > y_2)$.

8.4.3 Comparison methods and parameter instantiation

STUART was compared with representative time-to-event models: RSF [84], the lasso penalized Cox model (Lasso-Cox) [188], and Cox-SP-AFT which was explained in the end of Section 8.1. As a baseline model, we have reported the results of Cox regression with LASSO regularization. Lasso-Cox introduces the $L1$ norm penalty in the Cox log-likelihood function [188]. Since the majority of our used datasets are high-dimensional ($p \gg n$), we have employed Lasso-Cox due to its capability of handling high-dimensional datasets. For the purpose of estimating the generalization capacity of the models, a 5-fold cross-validation was performed on each dataset to determine test accuracy, and this process was repeated ten times to obtain reliable results. Throughout the ten iterations of the cross-validation process, the C-index is calculated for each test fold, and the result is the average value across the five folds. In Lasso-Cox, the optimal tuning parameter (λ) is selected by nested cross-validation, whereas no hyperparameter tuning has been applied to the other approaches. For RSF and STUART, the number of trees was set to 500, and the number of candidate variables considered in each tree node was set to $p/3$, where p is the number of variables.

8.5 Results and discussion

In this section, we first describe and compare the results of Lasso-Cox, Cox-SP-AFT, RSF, and STUART on the benchmark datasets described in Table 8.1. Then, we take a closer look at the results obtained for the NSBCD, Veteran,

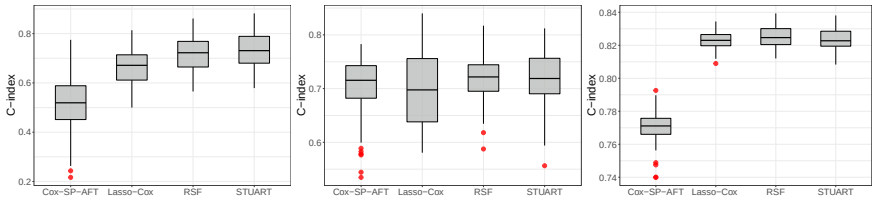
and NHANES I datasets as each represents a different type of dataset in terms of the number of features or observations.

Table 8.2 shows the means and standard deviations of the c-index on the datasets, as well as the average c-index of each algorithm. Based on the results, we can conclude that STUART is the winning approach, particularly in most high-dimensional datasets ($p \gg n$). More precisely, it can be concluded that on high dimensional datasets with a very small number of samples (e.g., Veer, LungBeer, AML, NSBCD, and GSE14764, all of which contain fewer than 120 observations), STUART is performing the best method. In addition, in several datasets with a high percentage of censored instances where very few labeled data are available (e.g., DBCD, GSE14764, EMTAB, and LungBeer, all with a higher than 72% censoring rate), STUART is a much better algorithm than RSF alone.

In this regard, despite the high censoring rates in both NHANES I and BRCA, RSF outperforms STUART, although only by a small margin given a large number of observations in both datasets.

Although STUART is the best algorithm for datasets with a high censoring rate and a small number of observations, at the other extreme, where the censoring rate is small (such as in Veteran and Lung), it does not perform as well. This may be the result of a lack of sufficient censored data to guide the augmentation process.

In light of the different behavioral patterns seen in different types of data, we selected three datasets that each represent a different type of dataset in terms of having either a very high number of features (NSBCD) or a very high number of observations (NHANES I) and a simple mid-size dataset (Veteran). The results on these datasets are illustrated by box plots in Figure 8.4. When comparing the range of C-indices (interquartile range), Cox-SP-AFT varies more dramatically and is the last algorithm in most experiments; but overall, STUART acts robustly and behaves like RSF. This robust behavior of STUART could be due to the fact that for censored instances, it compares the predicted survival time with the censoring time, which results in having more confident predictions. However, this should hold for Cox-SP-AFT as well since it also compares the predicted survival time with the censoring time. However, the reliability and stability of the Cox-SP-AFT model rely heavily on the accuracy of the AFT model and the single AFT model always encounters the robust issue in semi-supervised learning scenarios caused by heavy noise and even outliers [115, 22]. During the experiments, we also noticed that many censored data points always violate the constraint that losses should be smaller than (α), restricting the ability to fully exploit the censored data. Therefore, the AFT



(a) NSBCD (b) Veteran (c) NHANES I

Figure 8.4: Evaluation of the performance of the methods, for different percentages of labeled instances for three datasets.

model does not benefit from a large number of instances in order to be properly trained.

In comparison with the main competitor (RSF), STUART, although with a slight non-statistically significant margin, was ranked in a higher position according to the Friedman-Nemenyi test [1] presented in Figure 8.5 [41]. STUART outperforms Lasso-Cox and Cox-SP-AFT and manages to be statistically significantly better according to the Friedman-Nemenyi test. As a second-best method, RSF is statistically significantly superior to Cox-SP-AFT and has a slight non-significant lead over Lasso-Cox.

Overfitting is a concern that may arise when self-training approaches are used. Since our algorithms have no tunable hyperparameters, they are not prone to the kind of overfitting that results from the hyperparameter tuning process in other algorithms. Moreover, random forests overall are known to be robust to overfitting due to the fact that by increasing the number of trees, the variance of the error gets reduced.

Our findings demonstrate that the self-training technique that uses the information in the censored data points to guide the data augmentation process performs best, resulting in a competitive algorithm compared to RSF.

8.6 Conclusion

Predicting the time until an event of interest is a common problem encountered in medical applications, and it is traditionally addressed using survival analysis techniques. In this study, we have transformed the time-to-event prediction problem into a semi-supervised regression problem. Our findings indicate that

[1] In a critical distance diagram, those algorithms that are not joined by a line (i.e., their rankings differ more than a critical distance (CD)) can be regarded as statistically significantly different [41].

Table 8.2: Performance in terms of concordance index (C-index).

Datasets	Lasso-Cox	Cox-SP-AFT	RSF	STUART
Veteran	70.1 ± 6.5	70.05 ± 6.3	$\mathbf{71.5} \pm 5.3$	71.48 ± 5.4
Lung	$\mathbf{62.64} \pm 5.3$	60.82 ± 5.6	61.75 ± 5.1	62.01 ± 5.2
PBC	83.15 ± 3.5	80.51 ± 3.5	$\mathbf{83.5} \pm 3.1$	82.22 ± 4.5
DrAsGiven	52.53 ± 6.3	52.42 ± 10.5	57.42 ± 4.7	$\mathbf{57.74} \pm 7.7$
EMTAB386	51.43 ± 6.2	55.42 ± 8.9	50.14 ± 6.9	$\mathbf{59.11} \pm 5.9$
GSE14764	52.08 ± 8.5	54.63 ± 18.3	56.99 ± 17.3	$\mathbf{66.82} \pm 20.5$
GSE32062	52.11 ± 4.7	51.90 ± 7.3	50.03 ± 5.6	$\mathbf{56.12} \pm 6.4$
NSBCD	66.28 ± 8.2	51.05 ± 13.1	71.75 ± 6.5	$\mathbf{73.2} \pm 12.1$
Veer	62.63 ± 10.1	53.07 ± 10.2	67.4 ± 10.2	$\mathbf{71.71} \pm 11.6$
DBCD	68.99 ± 7.6	63.13 ± 7.8	73.5 ± 5.7	$\mathbf{74.15} \pm 6.1$
DLBCL	59.48 ± 6.4	55.74 ± 6.5	$\mathbf{59.7} \pm 4.4$	59.64 ± 5.9
LungBeer	50.93 ± 10.1	63.40 ± 14.3	67.55 ± 15.8	$\mathbf{72.34} \pm 11.3$
AML	55.90 ± 9.5	60.37 ± 8.9	60.02 ± 10.3	$\mathbf{64.99} \pm 3.2$
NHANES I	82.26 ± 0.52	77.06 ± 1.12	$\mathbf{82.5} \pm 0.6$	82.36 ± 0.6
BRCA	56.61 ± 6.3	56.22 ± 11.1	$\mathbf{63.62} \pm 5.2$	62.35 ± 5.3
Average	61.81	60.39	65.16	**67.75**

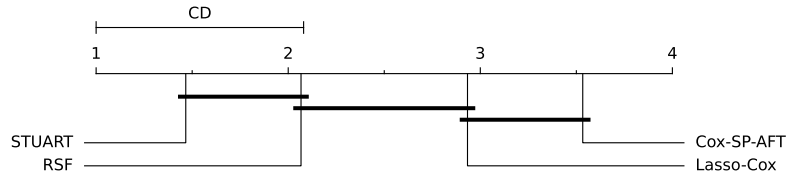


Figure 8.5: Results of the Friedman-Nemenyi test of methods ranking. The methods are compared in terms of their ranking using the evaluation measure, CI.

integrating partial supervision provided by censored data into a semi-supervised wrapper approach generally generates the best results. In this approach, called STUART, censored observations are introduced as partially labeled observations since their target values should exceed the censoring time. We have evaluated and compared the proposed approach on fifteen real-world survival analysis datasets, including clinical and high-dimensional ones. Our results have shown that our proposed approach especially in high-dimensional settings outperforms the others due to its ability to integrate partial supervision information inside a semi-supervised learning approach.

Further research can be carried out in several directions, of which we outline

a few below. In this study, we used STUART for survival analysis of right-censored data, but the same approach can be applied easily to left-censored data as well. The concept of idea that we proposed could be applied using other base learners and semi-supervised learning strategies, but it remains to be investigated whether the results carry over to other learners.

Declarations

Funding. This work was supported by KU Leuven Internal Funds (grant 3M180314).

Conflicts of interest/Competing interests. The authors have no conflicts of interest to declare that are relevant to the content of this article

Ethics approval. Not applicable

Consent to participate. No tests, measurements, or experiments were performed on humans as part of this work

Consent for publication. The authors have agreed to submit it in its current form for consideration for publication in Journal

Availability of data and material. All of the datasets used in this article are publicly available and have been referenced

Code availability. The source code will be publicly available

Authors' contributions.

Fateme Nateghi Haredasht: Investigation, Methodology, Software, Writing, Original draft.

Kazeem Adesina Dauda: Methodology, Software.

Celine Vens: Supervision, Conceptualization, Methodology, Review and editing

Acknowledgements

The authors also acknowledge the Flemish Government (AI Research Program).

Chapter 9

Predicting outcomes of acute kidney injury in critically ill patients using machine learning

The following chapter has been submitted to the Journal of the American Medical Informatics Association (JAMIA):

Nateghi Haredasht, F., Viaene, L., Pottel, H., De Corte, W., & Vens, C. Predicting outcomes of acute kidney injury in critically ill patients using machine learning.

Abstract

Background: Acute Kidney Injury (AKI) is a sudden episode of kidney failure that is frequently seen in critically ill patients. It has been shown that AKI is associated with chronic kidney disease (CKD) and mortality. We developed machine learning-based prediction models to predict outcomes following AKI stage 3 events in the intensive care unit.

Methods: We conducted a prospective observational study that used the medical records of ICU patients diagnosed with AKI stage 3. A random forest algorithm was used to develop two models that can predict patients who will progress to CKD after three and six months of experiencing AKI stage 3. Furthermore, two survival prediction models using random survival forests and survival XGBoost have been presented to predict mortality in these patients. We evaluated established CKD prediction models using cross-validation, receiver operating characteristics, and precision-recall curves and compared them with the baseline logistic regression models. An external test set has been used to evaluate the mortality prediction models, and the C-indices have been compared to the baseline COXPH model.

Result: We included 101 patients, of whom 75 and 53 made it to the first and second follow-ups which were three and six months after AKI development, respectively. The training set for the mortality prediction task has been increased by adding an additional unlabeled dataset consisting of 123 patients without information regarding their mortality. The RF (AUPR: 0.895 and 0.848) and XGBoost (c-index: 0.824) models have a better performance than the baseline models in predicting CKD and mortality, respectively.

Conclusion: The present work supports the view that machine learning-based models have the potential to advance clinical decision-making for identifying those patients that have a higher chance of developing CKD after hospital discharge from critically ill patients who experienced severe AKI. Also, we have shown that the inclusion of unlabeled data points in the survival analysis task results in achieving a better performance prediction.

9.1 Introduction

Acute kidney injury (AKI) is a very common complication in patients in the Intensive Care Unit (ICU), with up to 50% of patients having the condition [73]. A sudden increase in serum creatinine (SCr) and a decrease in urine volume are characteristic signs of this condition [97]. It is well established that AKI is strongly and independently related to short- and long-term outcomes, such as acute kidney disease (AKD), chronic kidney disease (CKD), and mortality [191]. Even though there have been considerable advances in the treatment of acute kidney injury, the outcomes, particularly in the severe stages, remain poor, with mortality levels often exceeding 50% and some survivors remaining dependent on renal replacement therapy (RRT) for a long period [191, 200, 40].

It has been reported that less than one-third of patients with AKI treated with renal replacement therapy (RRT) have their kidney function monitored by a nephrologist after surviving an episode of AKI [24]. The earlier CKD is diagnosed after AKI, though, the less intensive utilization of resources and the better the prevention of morbidity and mortality. Rather than following up with all AKI survivors after hospital discharge, it would be useful to identify subgroups of patients who are at higher risk for negative outcomes and follow them only. It is therefore necessary to develop prediction models in order to create a risk score at discharge time, that can be used to determine patient outcomes. Li et al. [112] developed prediction models for mortality in critically ill patients with AKI at 90 days and one year following the initiation of RRT. The study found that routinely collected features at the time of RRT initiation are limited in their ability to predict mortality among critically ill patients with AKI who are receiving RRT. In a separate study with a similar cohort, Järvisalo et al. [87] developed and validated new prediction models for ICU and hospital one-year mortality customized for patients with RRT-dependent AKI in which the developed models showed acceptable external validity in a validation cohort.

While some studies have proposed models that predict mortality after AKI, the prediction of other negative outcomes, such as CKD following AKI in critically ill patients, is an important but rather understudied area of research. Our literature study only yielded a study protocol for a multicenter prospective observational study, which predicts the occurrence of CKD at 3 years after patients suffered AKI during their ICU stay [54]. Possibly, the lack of such research is due to the fact that the follow-up of AKI survivors is considerably challenging after discharge from the ICU. The reason for this can be attributed to two primary factors. Firstly, the process is time-consuming and costly, and secondly, there is a high rate of dropouts. This can lead to reduced data collection and, in turn, can complicate subsequent analyses. This scarcity of labeled training data presents a key challenge in training supervised learning models in the biomedical field. There is, however, often plenty of unlabeled data

available for patients with similar characteristics and background information, e.g., from patients who fall outside the study period but otherwise fulfill all inclusion and exclusion criteria. In this article, we propose to take advantage of such unlabeled data by using it to expand the training set. In machine learning, the situation in which the training data consists of both labeled and unlabeled data is referred to as semi-supervised learning.

It is noteworthy that conventional statistical models, such as Logistic Regression, were the most used approaches in the literature to predict the outcomes of AKI. Nowadays, in the healthcare domain, machine learning algorithms have become increasingly popular due to their ability to handle large amounts of high-dimensional data, which is common in healthcare settings [42]. Electronic health records (EHRs) are capable of storing a large number of features and types, enabling accurate and reliable prediction models to be developed [88]. Furthermore, these machine learning models are capable of capturing complex interactions between the features in the datasets [7]. Despite this, there are still very few machine learning models that make use of the diversity and abundance of the data derived from EHRs.

Several machine learning models are currently being used to predict AKI [102, 106, 128, 48]. While a study developed predictive models for AKI stage 3 progression among critically ill patients who experienced AKI stage 1/2 using machine learning techniques [205], no such machine learning prediction models have been conducted for the prediction of CKD after experiencing AKI in the ICU.

The contribution of this study is fourfold. First, we conduct a follow-up study of critically ill patients who experienced AKI stage 3 during their ICU stay. Second, we employ machine learning-based prediction models to predict CKD after three and six months of developing AKI stage 3 in critically ill patients. Third, we employ machine learning-based time-to-event prediction models to predict mortality in critically ill patients who developed AKI stage 3. Finally, we examined whether we can obtain more accurate survival time predictions by using unlabeled data from patients who met the inclusion and exclusion criteria but were not included in the follow-up study.

The remainder of this paper is organized as follows: Section 9.2 presents a more detailed description of the study design, AKI definition, outcome of interests, dataset and the preprocessing methods, and the prediction methods; Section 9.3 presents the analysis results; Section 9.4 discusses and provides insights about our experiments; Finally, Section 9.5 brings our conclusions and future work directions.

9.2 Method

9.2.1 Study design

Two datasets were used in this study. First, a prospective observational study was conducted using ICU patients aged > 18 years who were diagnosed with AKI stage 3 during their stay in AZ Groeninge hospital in Kortrijk, Belgium, between September 2018 and October 2020. Exclusion criteria were patients with a baseline eGFR <30 ml/min/1.73 m^2 estimated by CKD-EPI [110], patients with renal replacement therapy (RRT) initiated before admission to the ICU, patients with a kidney transplant, patients with therapy restrictions with shift to palliative care, and patients who received extracorporeal blood purification techniques for reasons other than AKI. During the period of the patient's stay in the intensive care unit, data were collected using EHR. SCr and cystatin C (CysC) measurements were taken at the time of admission to the ICU and at the time of diagnosis of AKI (in most cases with a short time lag). The patients were also followed up by the nephrology department three, six, nine, and twelve months following diagnosis of AKI stage 3 in the intensive care unit. The eGFR was measured again during these follow-up visits.

In addition to the observational study, we have used an additional dataset containing patients who were not enrolled in this study and as a result, have not been followed up and have no information regarding outcomes. This dataset contains adult patients (> 18 years) who were diagnosed with severe AKI during their ICU stay at AZ Groeninge hospital in Kortrijk, Belgium, between January 2016 and September 2018 and between October 2020 and September 2021. The exclusion criteria for this dataset were identical to those used in the observational study.

For external validation of our mortality prediction model, we used data from patients who were admitted to the ICU and suffered severe AKI after the observational study ended (between October 2021 and June 2022).

9.2.2 Acute Kidney Injury classification

Patients with AKI stage 3, as defined by the KDIGO criteria, have been included in the study. KDIGO defines stage 3 as an increase in SCr up to 3 times from baseline within a 7-day period or urine output (UO) $< 0.3\ ml/kg/h$ for ≥ 24 hours [92]. In this study, true baseline SCr was available for patients who had an SCr measurement from an earlier visit (previously to their hospital or ICU admission). When such records were unavailable, baseline SCr was considered the first record of the patient's hospitalization before ICU admission. All SCr measurements were performed with an Enzymatic method that is traceable to the isotope dilution mass spectrometric method (IDMS), which

is the internationally approved reference method for measuring creatinine. In addition, CysC concentrations were measured by Liège University Hospital using a particle-enhanced nephelometric immunoassay on the BNII nephelometer (Siemens Healthcare Diagnostics, Marburg, Germany). The assay was calibrated against the internationally certified reference material ERM-DA471/IFCC for CysC.

9.2.3 Outcomes

The primary outcomes were the development of CKD three (and six) months after the AKI stage 3 event and mortality after the event. We defined CKD as eGFR $< 60mL/min/1.73m^2$, using the chronic kidney disease epidemiology collaboration (CKD-EPI) corresponding to CKD stage 3 or more according to the KDIGO classification.

9.2.4 Data description and preprocessing

During the ICU stay, demographic data (age, gender, weight, height, and BMI), comorbidity data, ICU interventions, the severity of illness scores (APACHE 2), the severity of disease classification (SAPS II), sequential organ failure assessment score (SOFA), admission diagnosis, events during ICU stay (respiratory, fluid balance, etc), laboratory data, and features concerning kidney function (SCr, CysC, eGFR, and albumin measurements) in which the latter was only collected for the purpose of CKD prediction, were reported. For some features like SOFA, SAPS II, and fluid balance, which were measured multiple times during the ICU stay, the average, the maximum, and the change (slope of the fitted line to that feature) have been calculated. As patients have been transferred to different types of ICUs based on their admission diagnosis, including Medical ICUs (MICUs), Surgical ICUs (SICUs), and Trauma Units, also this information was recorded. Comorbidities include a variety of features, such as arterial hypertension, chronic liver failure, and diabetes. ICU intervention includes the length of invasive ventilation (days), vasopressors, sedatives, antibiotics, and blood transfusions. In total 52 features (full feature set) have been used for the CKD prediction task. For the mortality prediction task, considering that informed consent was not obtained from the patients for the unlabeled data and their kidney function was not assessed in the same way, this dataset only consists of the features collected from the EHR excluding demographics and kidney-related features, resulting in 30 features (reduced feature set) in total[1].

[1]As the kidney related features were considered crucial for CKD prediction, we did not consider the unlabeled dataset in that prediction task.

Those features that were missing as a result of incomplete records were imputed by using the MICE method [193] for a maximum of 10 iterations.

9.2.5 Dataset characteristics

To predict mortality, the additional dataset has been used containing patients who were not enrolled in this study and as a result, have not been followed up and have no information regarding mortality. Due to the absence of information concerning the status of the mortality event (censored or observed), we refer to this dataset as the unlabeled data ($Udata$) with a corresponding status and time equal to zero. Also, for simplicity, we refer to the time-to-event data as labeled data ($Ldata$). This $Ldata$ consists of subjects who are either deceased (observed) or censored (alive at study end or at drop-out).

9.2.6 Prediction methods

A Random Forest (RF) [17] classifier has been employed to predict CKD after three and six months among patients who developed AKI stage 3 in the intensive care unit. RF is an ensemble learning method that constructs a multitude of decision trees (100 trees in our model) at the time of training and is used for classification, regression, and other tasks. Moreover, feature importance is calculated by weighting the mean decrease in impurity in splits with the given feature by the number of samples in the split [145].

A Random Forest (RF) [17] classifier has been employed to predict CKD after three and six months among patients who developed AKI stage 3 in the intensive care unit. RF is an ensemble learning method that constructs a multitude of decision trees (100 trees in our model) at the time of training and is used for classification, regression, and other tasks. Moreover, feature importance is calculated by weighting the mean decrease in impurity in splits with the given feature by the number of samples in the split [145].

The survival methods were trained on two scenarios of time-to-event data for mortality prediction. The first scenario consists of training the models on the resulting time-to-event data from the observational study (referred to as labeled data, or $Ldata$). This $Ldata$ consists of subjects who are either deceased (observed) or alive at the study end or at drop-out (censored). The second scenario consists of adding the unlabeled dataset (referred to as $Udata$) to the first scenario. Due to the absence of information concerning the status of the mortality event (censored or observed) in $Udata$, we set the corresponding status and time equal to zero [134].

We conducted a 5-fold cross-validation on the CKD prediction datasets in order to estimate the generalization capacity of the models. Results for the mortality

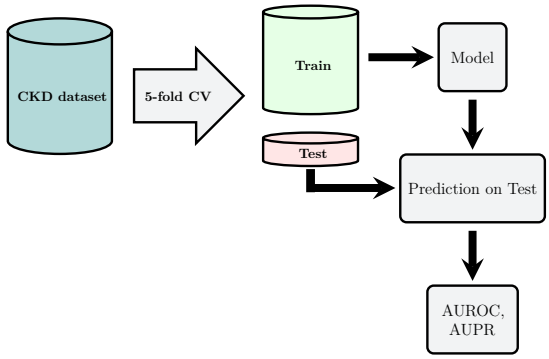


Figure 9.1: Study workflow for CKD prediction task. We utilized a population of patients from the observational follow-up data to train ML and statistical models to predict CKD after 3 and 6 months of developing AKI stage 3 in the ICU. 5-fold cross-validation was used to train and test models. Prediction performance was assessed with the AUROC and AUC-PR.

prediction task have been evaluated on the external dataset described in Section 9.2.1.

To better characterize the overall performance of the models, we contrasted the performance of ML algorithms against the conventional statistical methods Logistic Regression (LR) and Cox proportional hazards model (COXPH) [34] for CKD and mortality prediction, respectively.

The outline of the ML and statistical method workflow is shown for the CKD prediction and survival analysis in Figures 9.1 and 9.2, respectively. Figure 9.2 represents the two scenarios described earlier.

9.2.7 Statistical analysis

Categorical features were expressed as numbers (proportions) and continuous features as medians with interquartile ranges (IQR). For the CKD prediction task, the predictive performance of the models was compared using the area under the receiver-operating characteristic curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the net benefit using the decision curve analysis. AUC-ROC or ROC curves are constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at a variety of threshold settings. The true-positive rate is also known as sensitivity or recall, and the false-positive rate is known as (1 - specificity). The AUC-PR or PR curve illustrates the trade-off between Precision and Recall at various thresholds. In the context of decision curve analysis [197], the net benefit metric serves as a means of comparing the costs and benefits of various treatment approaches. Based on a model's sensitivity and specificity and the prevalence of the outcome

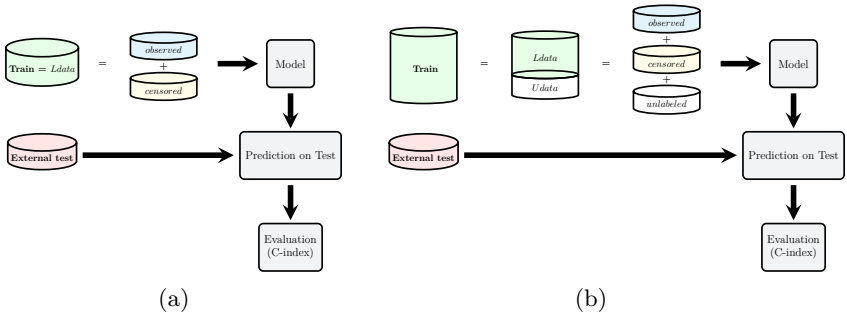


Figure 9.2: Study workflow for mortality prediction task (survival analysis). (a) In the first scenario, we utilized a population of patients from the observational follow-up data ($Ldata$) to train ML and statistical models to predict mortality in patients who developed AKI stage 3 in the ICU. In this scenario, the censoring rate is 57.42%. Prediction performance was assessed using C-index and has been tested on an external test set for each model separately. (b) In the second scenario, we utilized a population of patients from the observational follow-up data plus the unlabeled data ($Udata$) to have a bigger training set and train ML and statistical models to predict mortality in patients who developed AKI stage 3 in the ICU. In this scenario, the censoring rate is 80.8%. Prediction performance was assessed using C-index and has been tested on an external test set for each model separately.

in the population, the net benefit is calculated. According to decision curve analysis, the optimal strategy is the model with the highest utility. Through the cross-validation process, we calculated the predictive performance for each test fold and returned the average over the five test folds. For the mortality prediction task, Harrell's concordance index (C-index) [61] which is a common way to evaluate a model in survival analysis, has been used for comparing the predictive performance of the models. All analyses were performed using Python, version 3.9.

9.3 Results

9.3.1 Patient population

The study included 101 critically ill patients who developed AKI stage 3, with a median age of 74 (IQR 30-92) and 64 (63.3%) males. Characteristics of patients on ICU admission are shown in Table 9.1. Patients had two different measurements of both SCr and CysC: at the time of admission to the ICU, and at the time of developing AKI stage 3. Furthermore, patients who survived the ICU and were followed up successfully had two follow-ups (three and six months after the diagnosis of AKI). During the study period, 45% of the cohort (n=46) received dialysis for a median of 13 (1-160) days, with a mortality rate of 42.6% (n=43). There were 24 deaths during the ICU stay and 19 deaths

Table 9.1: Population Characteristics.

Characteristics	**N=101**
Demographics	
Age (years), median (IQR)	74 (30 - 92)
Female sex, %	36.6%
Body weight, kg	83 (45 - 150)
Body mass index, kg/m^2	27.7 (17 - 57)
Scores	
APACHE II	25.9 (11 - 43)
SOFA	9.2 (4.6 - 20.6)
SAPS II	55.2 (21.9 - 102.5)
ICU types	
MICU, %	80%
SICU, %	17%
Trauma, %	2%
Comorbidities	
Arterial hypertension, %	48%
Chronic Liver Failure	12%
Diabetes mellitus	22%
Chronic obstructive pulmonary disease	22%
Oncological history	22%
Suspected infection on admission	57%
ICU interventions	
Invasive ventilation days, median (IQR)	0.9 (0 - 20)
Fluid balance, median (IQR)	1032 (182 - 3470)
Transfusion, %	2.2%
Antibiotics, %	88%
Outcomes	
ICU days, median (IQR)	14.5 (1 - 160)
Hospital days, median (IQR)	26 (3 - 186)
CKD (after ICU discharge), %	49%
Mortality (hospital and follow-up), %	43%
Laboratory results	
Cystatin C (mg/L) at ICU admission	1.94 (0.67 - 8.06)
Creatinine (mg/dL) at ICU admission	1.98 (0.31 - 12.64)

during the follow-up phase, including two deaths between the time the patient was discharged from the ICU and the first follow-up. Patient dropouts and mortality resulted in a decrease in the number of patients attending follow-up appointments.

For the CKD prediction task, we used data from 101 patients with a full feature set resulting from the observational study. We have removed patients who deceased before three and six months follow-up after AKI, resulting in 75 and 53 patients in the training set, respectively. After three months, 47 of the 75 patients developed CKD, and 33 of the 53 patients developed CKD after six months.

For the mortality prediction task, we used $Ldata$ which has 101 subjects and $Udata$ with 123 subjects resulting in training the models with 224 subjects in total and with a reduced feature set. Moreover, the external validation set consists of 31 patients with the same reduced feature set.

9.3.2 Predictive performance: CKD prediction

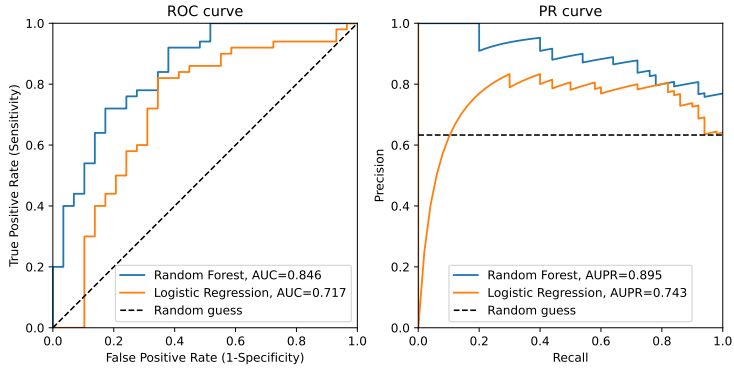
We performed a comparison between RF and LR for the prediction of CKD after three and six months of developing AKI stage 3, shown in Figure 9.3. In both tasks, RF had the highest predictive performance; however, with higher performance after 3 months (AUC 0.846; AUPR 0.895) compared to predicting CKD 6 months after AKI (AUC 0.803; AUPR 0.848). There is a reasonable explanation for this performance drop since fewer patients are involved in the second task (53 vs 75) and also, as time passes, patients recover and the ICU-related characteristics may have become less predictive of outcome.

In CKD prediction, we utilized the built-in algorithms for RF models to present feature importance.

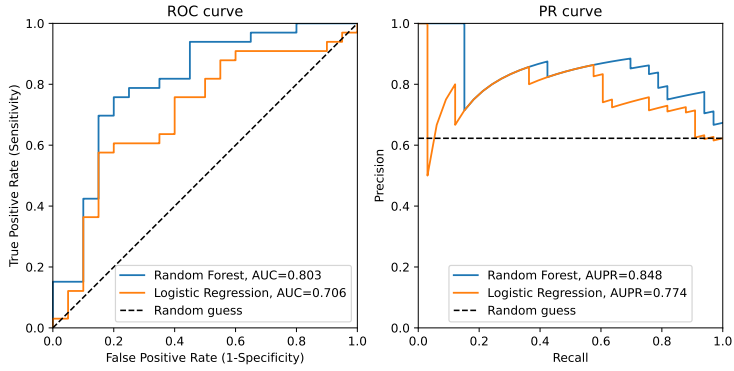
Figure 9.4 shows the top 20 features for CKD after AKI predicted by Random Forest models at 3 months and 6 months. According to both models, estimated GFR and creatinine value at the time of the AKI event are the most important features in the prediction of CKD. Besides other kidney-related features (absolute changes in SCr and CysC), the average fluid balance and SOFA score from patients' ICU stays are also predictive of the development of CKD three months after AKI. In addition, decision-curve analysis shows that compared to the reference model, the net benefit of RF models was larger over all the ranges of clinical threshold, indicating that the RF models prediction would more accurately identify high-risk patients (true-positives) while taking the trade-off with false-positives into consideration. However, the same conclusion cannot be made for LR models as in some thresholds ($\sim$ less than 40%) it is smaller than the clinical threshold (Figure 9.5).

9.3.3 Predictive performance: mortality prediction (survival analysis)

We also assessed survival prediction, in two different scenarios (Figure 9.2), made a comparison between three survival models, RSF, SGB, and COXPH, and reported the C-index performance on internal (Table 9.2) and external validation set (Table 9.3). The optimal model performance was the survival XGBoost model trained with a combination of labeled and unlabeled data, outperforming all other models with a C-index of 82%. Additionally, this XGBoost model demonstrated an increase of 3.4% in predictive performance compared to a model trained only using the labeled data. Although the RSF model also benefited by 1.42% from adding the unlabeled data to the training set, this increase was the highest for the COXPH model with 11.49%. As a result, our results also confirm that adding unlabeled instances to the training set improves the predictive performance on an independent test set [134].



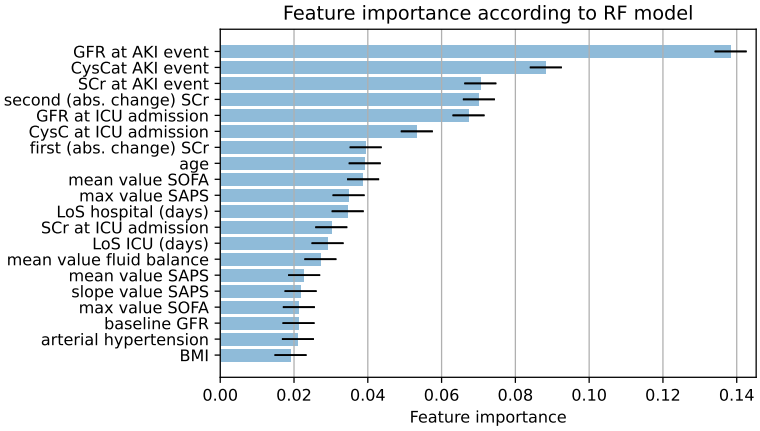
(a) Three months CKD prediction



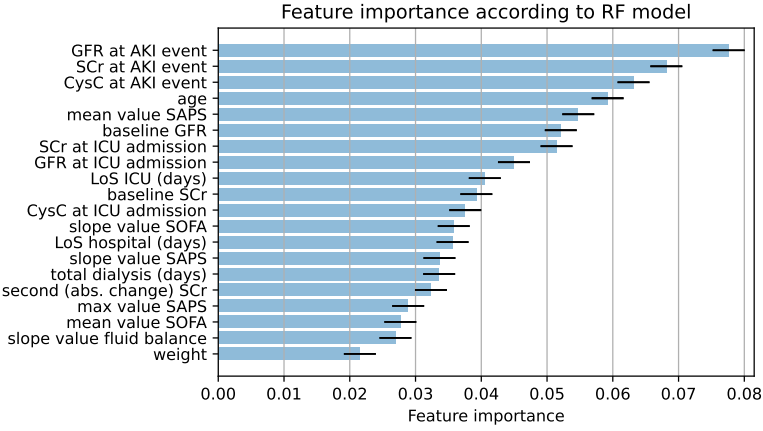
(b) Six months CKD prediction

Figure 9.3: Performance of the Random Forest and Logistic regression model. (a) Receiver operating characteristic and Precision–recall curves for estimating the discrimination between the Logistic regression model and the Random Forest model in the prediction of CKD three months after developing AKI. There are 75 subjects in this analysis from whom 63% developed CKD. (b) Receiver operating characteristic and Precision–recall curves for estimating the discrimination between the Logistic regression model and the Random Forest model in the prediction of CKD six months after developing AKI. There are 53 subjects in this analysis from whom 62% of them developed CKD.

Figure 9.6 shows the SHAP values of XGBoost (the best-performing survival model). For each feature, one point represents one patient. Positions along the x-axis represent the impact a feature has had on the model's output for that specific patient. Mathematically, this corresponds to the (logarithm of



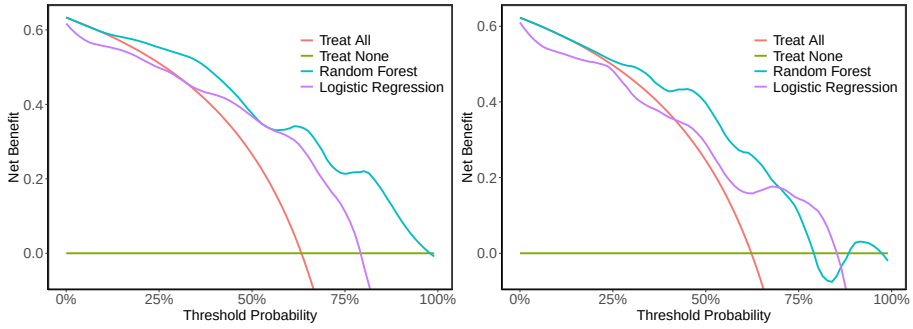
(a) Three months CKD prediction



(b) Six months CKD prediction

Figure 9.4: Feature importance for the top 20 features for CKD prediction after prediction. For each classifier, the feature importance estimation was based on mean decrease in impurity (MDI) calculations. In the features set, first (abs. change) SCr and second (abs. change) SCr show the absolute change in SCr at the ICU admission from baseline SCr and absolute change in SCr at the AKI event from baseline SCr, respectively.

the) mortality risk relative across patients (i.e., a patient with a higher SHAP value has a higher mortality risk relative to a patient with a lower SHAP value). On the y-axis, features are arranged in order of importance, which is determined by the mean of their absolute Shapley values. A feature's position in the plot indicates its importance for the model. According to the SHAP,



(a) Three months CKD prediction (b) Six months CKD prediction

Figure 9.5: Decision curve analysis graph showing the net benefit against threshold probabilities based on decisions from model outputs. The X-axis indicates the threshold probability for a positive CKD outcome; Y-axis indicates the net benefit.

Table 9.2: C-index performance on internal validation for mortality prediction.

	COXPH	RSF	XGBoost
Labeled	78.13	91.6	97.19
Labeled + Unlabeled	80.45	95.02	97.24

the average value and the evolution of severity of illness scores (SOFA and SAPS) during the ICU stay are important mortality risk factors for patients who experienced severe AKI. Among other features, the average amount of fluid balance contributes significantly to the prediction of mortality in AKI patients. However, the slope value for fluid balance seems to contribute adversely. A possible explanation could be that during the acute phase of critical illness, the fluid balance tends to be more positive after a few days (compared with the days before). Fluid balance "positivity" decreases as a patient's condition improves, eventually becoming negative. An intensivist usually begins diuretics or initiates dialysis when (s)he observes a daily increase in fluid balance positivity. It may lead to a more rapid return to normal fluid balance, or even a negative fluid balance (which is what diuretics are intended to accomplish).

9.4 Discussion

Due to the extensive use of EHRs as well as recent advances in machine learning, AI is expanding its influence on healthcare and has gradually changed the manner in which clinicians approach problem-solving [210]. However, to our best knowledge, there hasn't been any attempt to apply ML methods to predict the occurrence of CKD in AKI patients.

Table 9.3: C-index performance on external validation for mortality prediction.

	COXPH	RSF	XGBoost
Labeled	66.13	78.9	79.00
Labeled + Unlabeled	77.62	80.32	**82.48**

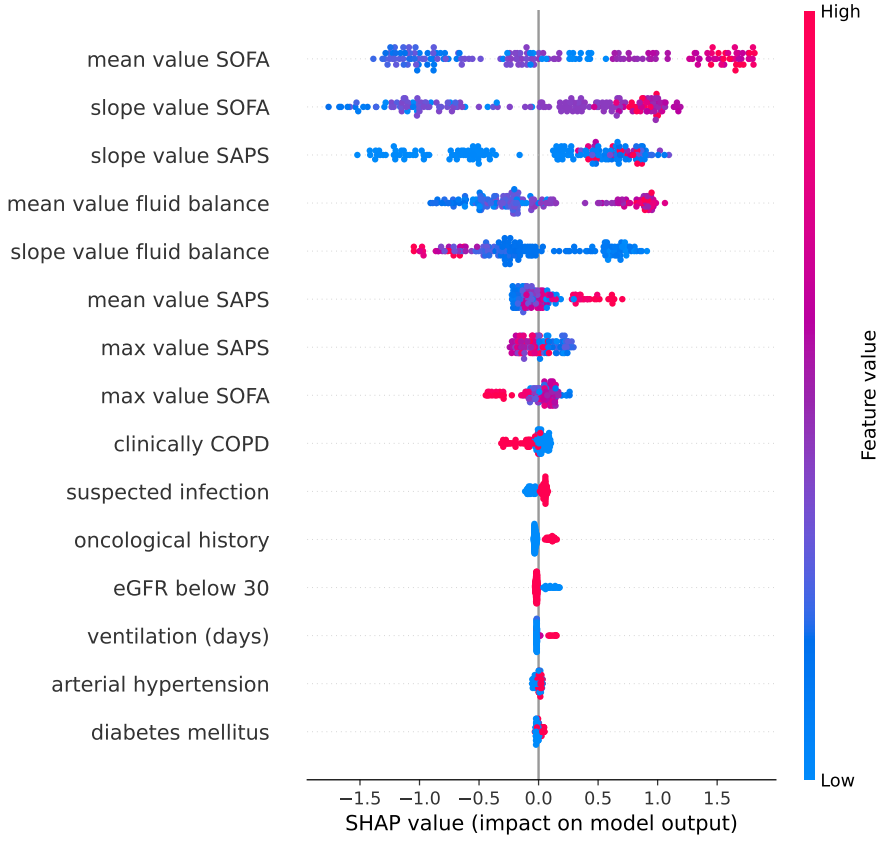


Figure 9.6: SHAP value of XGBoost model output. Each point represents a variable together with an observation. As demonstrated by the color bar, higher values are shown in red, while lower values are shown in blue.

The present study evaluated the potential utility of machine learning as a tool that can improve clinical decision-making. First, prediction models for CKD in critically ill patients after three and six months of experiencing severe AKI were explored using the random forest algorithm. In addition, as a time-to-event analysis, we have predicted mortality in the same group of patients

using two different survival models, namely random survival forest and survival XGBoost. In the CKD prediction task, the RF models had excellent performance with an AUPR of 0.895 and 0.848 for three months and six months CKD, respectively, which was significantly better than the performance of the baseline logistic regression model (0.743 and 0.774, respectively). The results of variable importance ranking can also potentially inform clinical practice. In our analysis, the importance of creatinine and cystatin C at the time of developing AKI was determined by interpreting the importance of each variable in the RF models to predict CKD. In addition, the decision curve analysis indicates that these models show a benefit compared to the 'treat everyone' and 'treat none' approaches, which indicates the possibility of allocating targeted assessments and interventions in addition to those broad health service strategies.

Overall, our results in predicting CKD after severe AKI showed that machine learning models are tools that can be helpful in clinical decision-making. The presented ML models can be quite useful in practice since they are based on features that are routinely collected in ICU and laboratory data, which are easily assessable in most ICU units. In addition, the need for early detection and prevention of CKD is important. However, currently, after discharge from ICU, the follow-up of AKI survivors is considerably challenging mainly due to being time-consuming and costly, and patients drop out. As a result, using the presented ML models, a risk profile can be developed for each survivor of AKI using electronic health records (EHR) data upon discharge from the ICU, which allows clinicians to create a customized follow-up plan.

Moreover, in our second objective, the survival XGBoost model had a significantly higher performance with a c-index of 0.79 in predicting mortality compared to the baseline COXPH model with a c-index of 0.661. The result for the XGBoost model even improved to a c-index of 0.824 when adding a set of baseline data that has no information about the mortality (unlabeled data) to the training set. Follow-up studies in clinical settings often experience a reduction in data collection, which can complicate subsequent analyses. On the other hand, there is often a substantial amount of baseline data available on patients with similar characteristics and background information, for example, from patients outside the study time window. Our analysis has shown that such unlabeled data instances can be used to predict survival times with a high degree of accuracy.

In light of these findings, ML seems to be a viable approach to predicting CKD progression and mortality in AKI patients, which may help physicians establish personalized treatment plans at an early stage.

The study we conducted has a number of strengths. First, we examined not only discrimination but also clinical usefulness, which was considered a key

measure of the performance of the model. In addition, we also validated the prediction model for our primary outcome of mortality in a distinct cohort of critically ill patients who developed AKI. Furthermore, the explanations and interpretations that accompany the predicted chance of CKD and the overall survival are intended to give clinicians insights regarding how each prognostic factor contributes to CKD development and overall survival.

There are some limitations to our study. First, this cohort consisted of a few subjects which might have affected model performance as discussed earlier. Also, a single center provided the data for our models, which may have compromised their generalizability.

9.5 Conclusion

According to our study, machine learning models are able to improve the prediction of CKD and mortality in critically ill patients who developed severe AKI during their stay in the intensive care unit. Our presented models had excellent performance in predicting CKD and mortality and were significantly better than the performance of the baseline models. In addition, we have investigated the inclusion of unlabeled data points in the survival analysis task. More precisely, we have shown that learning from data with three degrees of supervision: fully observed, partially observed (censored), and unobserved (unlabeled) data points lead to better performance in mortality prediction. Although our time-to-event models have been tested on an external validation set, the CKD prediction models need to be tested externally.

Declarations

9.6 Ethics approval and consent to participate

Funding. This work was supported by KU Leuven Internal Funds (grant 3M180314).

Conflicts of interest/Competing interests. The authors have no conflicts of interest to declare that are relevant to the content of this article.

Consent for publication. The authors have agreed to submit it in its current form for consideration for publication in Journal.

Acknowledgements

The authors also acknowledge the Flemish Government (AI Research Program).

Chapter 10

General discussion

Critically ill patients are admitted to the intensive care unit (ICU) with acute and life-threatening conditions. Approximately 40% of critically ill patients are affected by Acute Kidney Injury (AKI). Different studies have linked AKI to the development of chronic kidney disease, end-stage kidney disease, and mortality, suggesting that even a short episode of acute kidney injury might lead to long-term morbidity and mortality. Early detection and treatment can often keep chronic kidney disease from leading to kidney failure, which requires dialysis or a kidney transplant to maintain life. However, currently, follow-up of AKI survivors is often lacking and not standardized: follow-up of kidney function by a nephrologist in patients surviving an episode of AKI treated with RRT is reported in approximately one-third of the patients. Close follow-up and interventions aimed at preserving kidney function may positively impact long-term outcomes. However, the follow-up of AKI survivors is considerably challenging after discharge from the ICU. The reason for this can be attributed to two primary factors. Firstly, the process is time-consuming and costly, and secondly, there is a high rate of dropouts. To this end, novel machine learning models to predict the risk of chronic kidney disease for AKI survivors are certainly needed.

Survival analysis is an important part of medical research, which is often used for determining prognostic indices for mortality and recurrence of diseases, and for analyzing treatment outcomes. Surprisingly, in the data mining or machine learning research community, survival data has not received much attention. Standard machine learning techniques cannot be straightforwardly applied to survival data, mainly because of the censoring. Censoring is a form of missing data problem in which time to event is not observed. Nevertheless, a number of studies transform the survival data into a format suitable for standard machine

learning techniques, inevitably losing information. Often, the task becomes a binary classification task (does the patient survive a particular time point?), and censored data points are either removed, or their impact is decreased through a weighting technique. Several studies use semi-supervised learning techniques to deal with censored survival data. They treat the censored data points as unlabeled, thereby ignoring the survival information that they represent. Again, they often convert the survival prediction problem into a binary classification problem.

The objective of this thesis was to develop and validate decision-support applications to assist clinicians in managing critically ill patients who developed AKI. An important contribution of this Ph.D. project was the adoption of existing and the development of new machine learning techniques. The project objective could be separated into methodological objectives, which make a substantial contribution to the field of machine learning, and medical application objectives, which will lead to an improved post-ICU policy for AKI patients. Below, we explain the contributions of the conducted research.

10.1 Discussion on results and contributions

The first steps in this PhD, presented in Chapter 4, consisted of investigating different AKI definitions and their association with clinical outcomes. Due to the existence of different AKI definitions, analyzing AKI incidence and associated outcomes is challenging. We investigated the incidence of AKI events defined by 4 different definitions (standard AKIN and KDIGO, and modified AKIN-4 and KDIGO-4) and its association with in-hospital mortality. A multivariate logistic regression analysis was used to determine if there was an association between AKI stages and in-hospital mortality. A classification task was also conducted using two prediction models (random forest and logistic regression) to predict mortality in AKI patients defined by these four definitions. Results showed that KDIGO-4 is more sensitive in detecting AKI events. In-hospital mortality increased as the stage of AKI events increased for both KDIGO-4 and AKIN-4; however, KDIGO-4 (KDIGO) has a higher odds ratio at a higher stage of AKI compared to AKIN-4 (AKIN). Lastly, we confirmed that within the KDIGO AKI stage 1, there are two subpopulations with different severity of clinical outcomes (mortality).

In order to thoroughly review the existing validated risk prediction models for developing poor renal outcomes after AKI scenarios, in Chapter 5, a systematic review has been performed. Medline, EMBASE, Cochrane, and Web of Science were searched for articles that developed or validated a prediction model. In total, nine articles met the inclusion criteria from which one study was a study protocol, resulting in eight final studies. After reviewing the found articles, we concluded

that risk prediction models for developing renal insufficiency after experiencing AKI are based on simple statistical/machine learning models. We believe that advanced machine learning models using big data information are required to increase the predictive performance for developing renal insufficiencies.

It is well known that serum creatinine (SCr), one of the most widely used biomarkers to assess kidney function, does not always accurately predict glomerular filtration rate (GFR) due to its dependence on non-GFR determinants, such as muscle mass and recent meat consumption. Cystatin C (CysC), another biomarker of kidney function, has attracted the attention of researchers and clinicians. In Chapter 6, we compared GFR estimation using SCr and CysC in detecting CKD over a 1-year follow-up after an AKI-stage 3 event in the ICU, as well as analyzed the association between eGFR (using SCr and CysC) and mortality after the AKI event. Our results demonstrated that the incidence of CKD was highly discrepant with $\text{eGFR}_{\text{CysC}}$ versus eGFR_{SCr} during the follow-up period. CysC detects more CKD events compared to SCr in the follow-up phase and $\text{eGFR}_{\text{CysC}}$ is a predictor for mortality in follow-up but not eGFR_{SCr}. Determining the proper marker to estimate GFR in the post-ICU period in AKI stage 3 populations needs further study to improve risk stratification.

Many clinical time-to-event studies that require follow-up of the patients after a hospital stay form a logistic challenge because once the patients take up their normal activities, it is often difficult to reach or motivate them to continue their participation in the study. Thus, drop-out is frequently observed, and for many patients, no follow-up data is available at all (only data from their hospital stay). In addition to this, for many of these prospective studies, the training set can be easily augmented with retrospective hospital data from patients not participating in the study. If the study outcome is determined during follow-up, for both scenarios, this means that we often have a considerable unlabeled part of the training set (equivalently, the censoring time is zero for these patients). Based on successes in the semi-supervised learning domain, we showed that unlabeled data, which is often easy to obtain, can increase the predictive performance in a survival prediction task. In Chapter 7, we proposed three approaches to deal with this novel setting and provide an empirical comparison over fifteen real-life clinical and gene expression survival datasets. Our results demonstrated that all approaches are able to increase the predictive performance over independent test data. We also showed that integrating the partial supervision provided by censored data in a semi-supervised wrapper approach generally provides the best results, often achieving high improvements, compared to not using unlabeled data.

Inspired by our proposed method in Chapter 7, we developed a semi-supervised-based model to predict time-to-event. In Chapter 8, we used a self-training

wrapper approach with random survival forests as the base learner in which censored observations were introduced as partially labeled observations since their predicted time (target value) should exceed the censoring time. We concluded that the ability of our proposed approach to integrate partial supervision information within a semi-supervised learning strategy has enabled it to achieve competitive performance compared to baseline models, particularly in the case of a high-dimensional regime.

Next, Chapter 9 presents our ICU application where we addressed predicting outcomes following AKI stage 3 events in the intensive care unit. While AKI has a high incidence rate in ICU patients and is associated with considerable societal and economic consequences as previously mentioned, once these patients leave ICU (and hospital), they often disappear from nephrologist follow-up if they are no longer dialysis dependent. Thus, it is important to accurately estimate the risk of progression for these patients, in order to develop a rational policy with respect to medical care and follow-up. A random forest algorithm was used to develop two models that can predict patients who will progress to CKD after experiencing AKI stage 3. Furthermore, two survival prediction models using random survival forests and survival XGBoost have been presented to predict mortality in these patients. According to our study, machine learning models were able to improve over the classical statistical techniques in the prediction of CKD and mortality in critically ill patients who developed severe AKI during their stay in the intensive care unit. Our presented models had good performance in predicting CKD and mortality and were significantly better than the performance of the baseline models. In addition, we have investigated the inclusion of unlabeled data points presented in Chapter 7 in the survival analysis task. We showed that adding the unlabeled data points to the training set led to better performance in mortality prediction.

10.2 Conclusion

In this thesis, with extensive collaboration between clinical experts and computer scientists, we were able to gain novel insights into critical care medicine through the use of machine learning methods. First, we demonstrated that we could predict accurate survival times by including baseline data that are available for patients with similar characteristics and background information to the available dataset. Second, we developed and validated novel machine-learning-based time-to-event prediction models, which achieved competitive performance compared to baseline models, particularly in the case of a high-dimensional regime.

In addition, in the context of ICU applications, we have tackled the task of accurately estimating the risk of CKD progression for AKI patients, to develop

a rational policy concerning medical care and follow-up. Interpretable machine learning models, including decision tree-based models, were our main focus since interpretability is especially important in healthcare.

Finally, we hope that the findings obtained in the examined domains of critical illness will open up avenues for prospective interventions to evaluate whether the tools developed or the novel insights can improve the process of care and patient outcomes.

10.3 Future research direction

The work presented in this thesis can be extended in many different ways. Described below are the most promising directions for future research.

In Chapters 7 and 8, we have employed random survival forest as the base learner in the self-training approach. However, in principle, the basic ideas of our approach are transferable to other base learners. An interesting topic for future work could be to investigate whether changing the base learner would increase the predictive performance of the presented models.

Also, survival analysis tasks usually assume a single event per data sample. In practice, many studies are interested in predicting the time to multiple events, or to recurrent events. Instead of breaking up the prediction task into multiple tasks (e.g., one per event), developing a multi-output survival analysis tool, based on the machine learning literature on multi-output prediction could be a potential next step. In the context of AKI, time to multiple events could be time to develop new episodes of AKI and CKD.

Another interesting extension could be the application of deep learning in survival analysis. Recently, two studies have applied Transformers to the field of survival analysis [203, 81]. Transformers are deep learning models that adopt self-attention by differentially weighting the significance of each part of the input data.

In addition, the databases of electronic health records contain patient health information in the form of tabular data, which refers to data that has been organized in a table with rows and columns. One idea can be to benefit from medical image data by extracting some informative radiomics features using deep learning-based models or benefit from free-text clinical notes and then combining them with the EHR data in order to further improve the prediction performance.

Appendix A

Appendix

The content presented below is related to the experiments presented in Chapter 4.

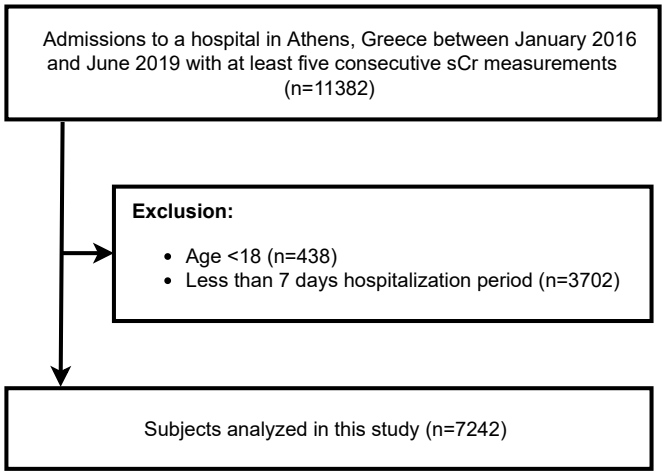


Figure A.1: Flowchart for patient inclusion.

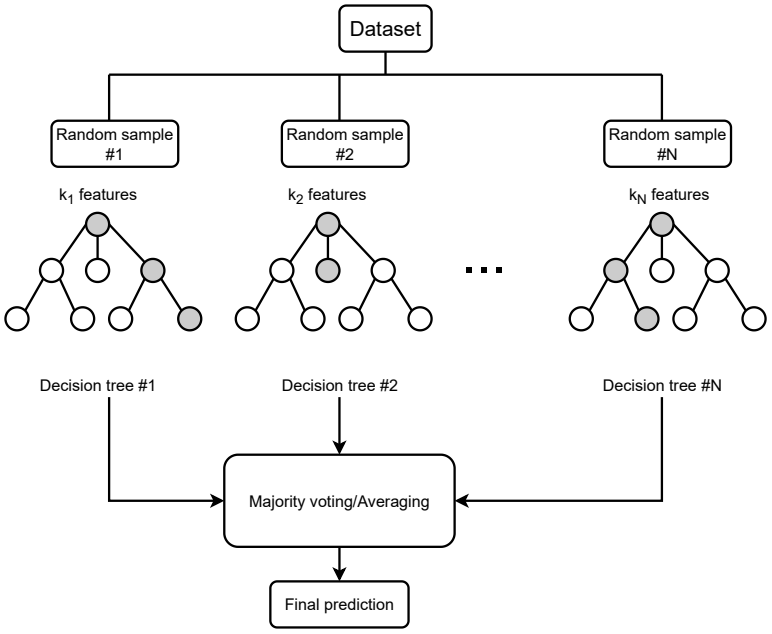


Figure A.2: Example of an ensemble of decision trees (random forest).

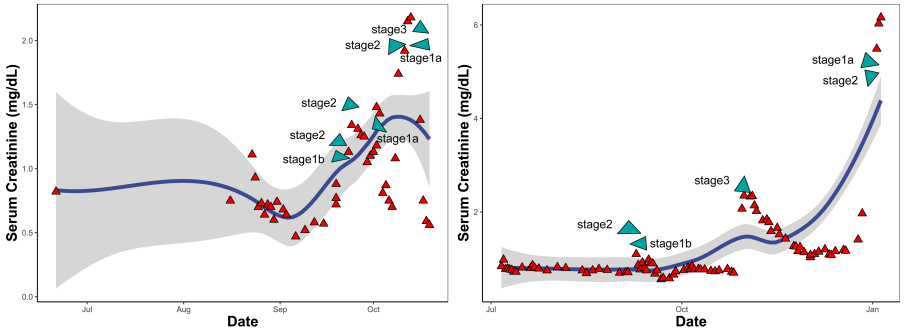


Figure A.3: Examples of SCr-time trajectories for two random patients using KDIGO-4.

Table A.1: Overview of patients with and without AKI-events, according to AKIN-4 and KDIGO-4 definitions.

Patients		AKIN-4		
		No AKI	AKI	Total
KDIGO-4	No AKI	5711	2	5713
	AKI	461	1068	1529
	Total	6172	1070	7242

Table A.2: Overview of deaths with and without AKI-events, according to AKIN-4 and KDIGO-4 definitions.

Deaths		AKIN-4		
		No AKI	AKI	Total
KDIGO-4	No AKI	145	0	145
	AKI	101	443	1529
	Total	246	443	689

Table A.3: Counts of patients experiencing only stage 1a, stage 1b, and both stage 1a and stage 1b.

	Only Stage 1a	Mortality (%)	Only Stage 1b	Mortality (%)	Stage 1a and stage 1b	Mortality (%)
AKIN-4	544	176 out of 544 (32.35%)	201	65 out of 201 (32.33%)	176	121 out of 176 (68.75%)
KDIGO-4	222	28 out of 222 (12.61%)	557	115 out of 557 (20.65%)	119	51 out of 119 (42.86%)

Table A.4: Incidence of AKI in terms of most severe case and mortality rate.

	Stage 1a	Mortality (%)	Stage 1b	Mortality (%)	Stage 2	Mortality (%)	Stage 3	Mortality	Total cases
AKIN-4	544	176 out of 544 (32.35%)	377	186 out of 377 (49.34%)	125	66 out of 125 (52.8%)	24	15 out of 24 (62.5%)	1070
KDIGO-4	222	28 out of 222 (12.61%)	676	166 out of 676 (24.55%)	416	211 out of 416 (50.72%)	215	139 out of 215 (64.65%)	1529

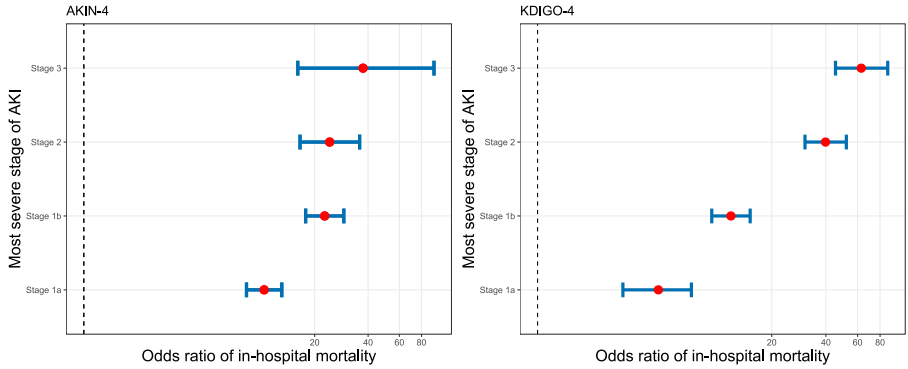


Figure A.4: Odds ratio of in-hospital mortality using logistic regression (the blue bars are the 95%CI), stratified by the most severe stage of AKI-events according to AKIN-4 and KDIGO-4 definitions for AKI-events.

Table A.5: Incidence of AKI in terms of most severe case and mortality rate.

	Random Forest					**Logistic Regression**				
	Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
KDIGO-4	0.838	0.385	0.766	0.512	0.833	0.838	0.385	0.766	0.512	0.856
AKIN-4	0.869	0.441	0.657	0.528	0.779	0.869	0.441	0.664	0.530	0.796
KDIGO	0.837	0.383	0.766	0.511	0.825	0.837	0.383	0.766	0.511	0.854
AKIN	0.881	0.476	0.642	0.546	0.770	0.881	0.476	0.642	0.546	0.802

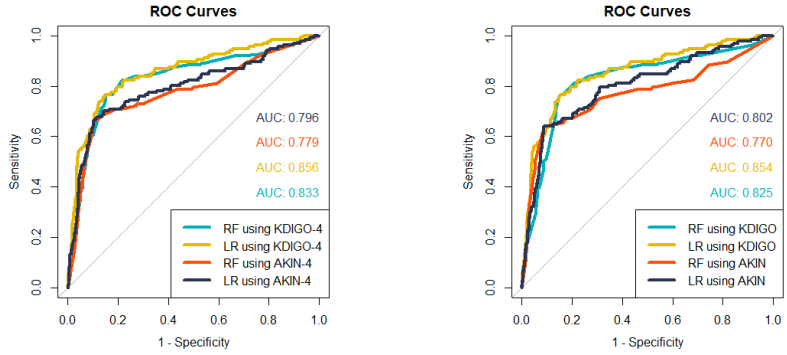


Figure A.5: ROC curves for RF and LR using KDIGO-4 and AKIN-4 (left) and using KDIGO and AKIN (right).

The content presented below is related to the experiments presented in Chapter 6.

Table A.6: eGFR (mL/min/1.73 m^2) statistics by using biomarkers in ICU and follow-up phase.

	ICU stay				Follow-up phase							
	ICU admission		AKI diagnosis		1st visit		2nd visit		3rd visit		4th visit	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
CysC	37.2	27.9	23.3	21	37	16.3	43.5	19.1	45.7	19.6	51.8	19.2
SCr	32.2	49.7	18.5	34	50.2	39.4	55.8	27.9	52.5	40.2	67.4	34.1

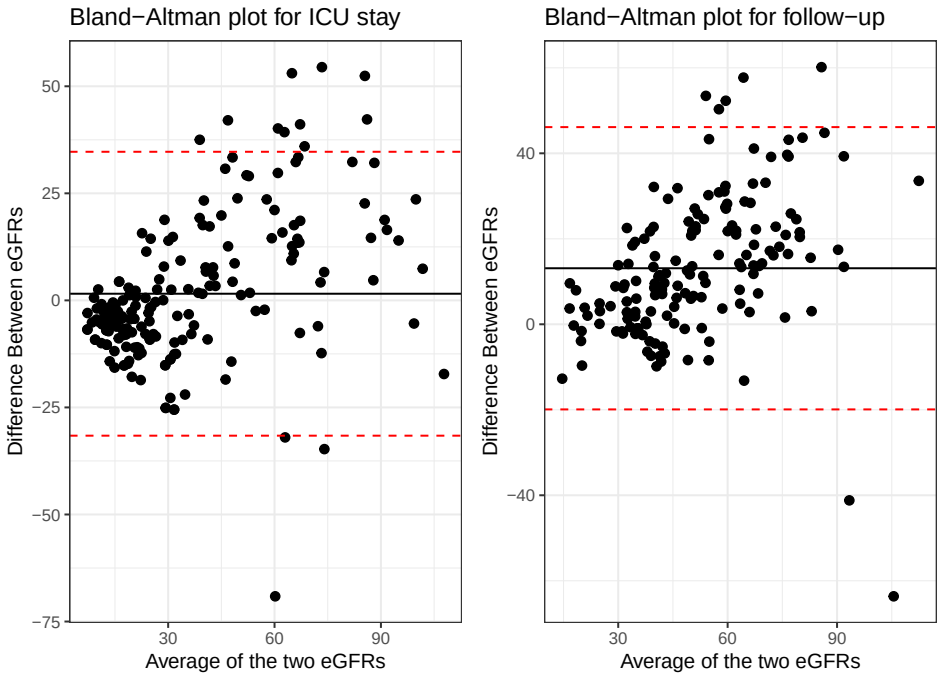


Figure A.6: Bland-Altman analysis of the two eGFRs during ICU stay and follow-up phase.

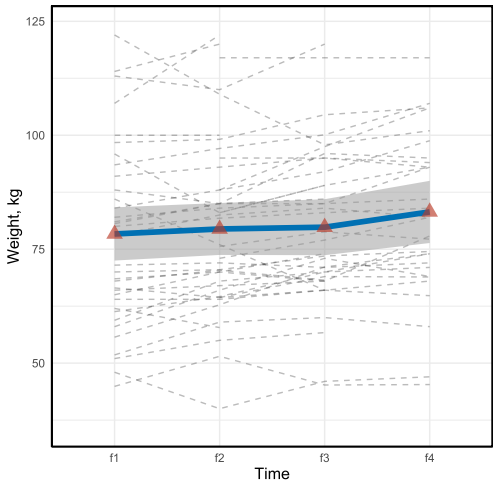


Figure A.7: Individual trajectories for weight during the follow-up. The dashed gray lines represent each subject, the red triangles show the average weight values at that specific time point, and the blue lines are smooth curves obtained via LOESS.

Table A.7: Number of patients with eGFR $<$ 60 mL/min/1.73 m^2 and eGFR $\geq$ 60 mL/min/1.73 m^2 based on SCr and CysC in the 1st follow-up visit.

CKD		$eGFR_{SCr}$		Total
		<60	≥ 60	
$eGFR_{CysC}$	<60	33	19	52
	≥ 60	1	7	8
Total		34	26	60

Table A.8: Overview of patients with CKD stages according to eGFR using SCr and CysC during the 1st follow-up.

CKD		$eGFR_{SCr}$						
		CKD1	CKD2	CKD3A	CKD3B	CKD4	CKD5	Total
$eGFR_{CysC}$	CKD1	1	0	0	0	0	0	1
	CKD2	3	3	1	0	0	0	7
	CKD3A	2	4	1	0	0	0	7
	CKD3B	1	11	6	9	1	0	28
	CKD4	0	1	1	8	4	1	15
	CKD5	0	0	0	0	2	0	2
	Total	7	19	9	17	7	1	60

Table A.9: Overview of patients with CKD stages according to eGFR using SCr and CysC during the 2nd follow-up.

CKD		$eGFR_{SCr}$						
		CKD1	CKD2	CKD3A	CKD3B	CKD4	CKD5	Total
$eGFR_{CysC}$	CKD1	0	1	0	0	0	0	1
	CKD2	1	3	0	0	0	0	4
	CKD3A	3	5	4	0	0	0	12
	CKD3B	0	3	7	5	0	0	15
	CKD4	0	0	3	2	1	0	7
	CKD5	0	0	0	0	1	0	1
	Total	4	12	14	7	2	0	39

Table A.10: Overview of patients with CKD stages according to eGFR using SCr and CysC during the 3rd follow-up.

CKD		$eGFR_{SCr}$						
		CKD1	CKD2	CKD3A	CKD3B	CKD4	CKD5	Total
$eGFR_{CysC}$	CKD1	0	0	0	0	0	0	0
	CKD2	3	2	0	0	0	0	5
	CKD3A	1	7	3	2	0	0	13
	CKD3B	0	1	4	6	0	0	11
	CKD4	0	0	1	1	3	0	5
	CKD5	0	0	0	0	0	0	0
	Total	4	10	9	8	3	0	34

Table A.11: Overview of patients with CKD stages according to eGFR using SCr and CysC during the 4th follow-up.

CKD		$eGFR_{SCR}$						
		CKD1	CKD2	CKD3A	CKD3B	CKD4	CKD5	Total
$eGFR_{CysC}$	CKD1	0	1	0	0	0	0	1
	CKD2	2	7	0	0	0	0	9
	CKD3A	0	4	2	3	0	0	9
	CKD3B	0	1	0	4	0	0	5
	CKD4	0	0	0	0	1	0	1
	CKD5	0	0	0	0	0	0	0
	Total	2	13	2	7	1	0	25

Bibliography

[1] Adhikari, N. K., Fowler, R. A., Bhagwanjee, S., and Rubenfeld, G. D. Critical care and the global burden of critical illness in adults. *The Lancet 376*, 9749 (2010), 1339–1346.

[2] Ahmad, M. A., Eckert, C., and Teredesai, A. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (2018), pp. 559–560.

[3] Awad, A., Bader-El-Den, M., McNicholas, J., and Briggs, J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International journal of medical informatics 108* (2017), 185–195.

[4] Bagshaw, S. M., George, C., Bellomo, R., Committee, A. D. M., et al. Early acute kidney injury and sepsis: a multicentre evaluation. *Critical Care 12*, 2 (2008), R47.

[5] Bair, E., and Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol 2*, 4 (2004), e108.

[6] Ballinger, B., Hsieh, J., Singh, A., Sohoni, N., Wang, J., Tison, G. H., Marcus, G. M., Sanchez, J. M., Maguire, C., Olgin, J. E., et al. Deepheart: semi-supervised sequence learning for cardiovascular risk prediction. *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[7] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence 41*, 2 (2018), 423–443.

[8] Beker, B. M., Corleto, M. G., Fieiras, C., and Musso, C. G. Novel acute kidney injury biomarkers: their characteristics, utility and concerns. *International urology and nephrology 50*, 4 (2018), 705–713.

[9] Bellomo, R., Kellum, J. A., and Ronco, C. Acute kidney injury. *The Lancet 380*, 9843 (2012), 756–766.

[10] Bellomo, R., Kellum, J. A., Ronco, C., Wald, R., Martensson, J., Maiden, M., Bagshaw, S. M., Glassford, N. J., Lankadeva, Y., Vaara, S. T., et al. Acute kidney injury in sepsis. *Intensive care medicine 43*, 6 (2017), 816–828.

[11] Bellomo, R., Ronco, C., Kellum, J. A., Mehta, R. L., and Palevsky, P. Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: the second international consensus conference of the acute dialysis quality initiative (adqi) group. *Critical care 8*, 4 (2004), 1–9.

[12] Bendavid, I., Statlender, L., Shvartser, L., Teppler, S., Azullay, R., Sapir, R., and Singer, P. A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from covid-19. *Scientific reports 12*, 1 (2022), 1–14.

[13] Bishop, C. M., and Nasrabadi, N. M. *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[14] Blockeel, H., Raedt, L. D., and Ramon, J. Top-down induction of clustering trees. In *Proceedings of the Fifteenth International Conference on Machine Learning* (San Francisco, CA, USA, 1998), ICML '98, Morgan Kaufmann Publishers Inc., pp. 55–63.

[15] Boonstra, A., Versluis, A., and Vos, J. F. Implementing electronic health records in hospitals: a systematic literature review. *BMC health services research 14*, 1 (2014), 370.

[16] Bouchard, J., and Mehta, R. L. Acute kidney injury in western countries. *Kidney Diseases 2*, 3 (2016), 103–110.

[17] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[18] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and regression trees*. Routledge, 2017.

[19] Callahan, A., and Shah, N. H. Machine learning in healthcare. In *Key Advances in Clinical Informatics*. Elsevier, 2017, pp. 279–291.

[20] Case, J., Khan, S., Khalid, R., and Khan, A. Epidemiology of acute kidney injury in the intensive care unit. *Critical care research and practice 2013* (2013).

[21] Cerdá, J., Mohan, S., Garcia-Garcia, G., Jha, V., Samavedam, S., Gowrishankar, S., Bagga, A., Chakravarthi, R., Mehta, R., Group, C., et al. Acute kidney injury recognition in low-and middle-income countries. *Kidney international reports 2*, 4 (2017), 530–543.

[22] Chai, H., Li, Z.-n., Meng, D.-y., Xia, L.-y., and Liang, Y. A new semi-supervised learning model combined with cox and sp-aft models in cancer survival analysis. *Scientific reports 7*, 1 (2017), 1–12.

[23] Chapelle, O., Scholkopf, B., and Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks 20*, 3 (2009), 542–542.

[24] Chawla, L. S., Amdur, R. L., Amodeo, S., Kimmel, P. L., and Palant, C. E. The severity of acute kidney injury predicts progression to chronic kidney disease. *Kidney international 79*, 12 (2011), 1361–1369.

[25] Chawla, L. S., Bellomo, R., Bihorac, A., Goldstein, S. L., Siew, E. D., Bagshaw, S. M., Bittleman, D., Cruz, D., Endre, Z., Fitzgerald, R. L., et al. Acute kidney disease and renal recovery: consensus report of the acute disease quality initiative (adqi) 16 workgroup. *Nature Reviews Nephrology 13*, 4 (2017), 241–257.

[26] Chawla, L. S., and Kimmel, P. L. Acute kidney injury and chronic kidney disease: an integrated clinical syndrome. *Kidney international 82*, 5 (2012), 516–524.

[27] Chen, Y., Zelnick, L. R., Wang, K., Katz, R., Hoofnagle, A. N., Becker, J. O., Hsu, C.-Y., Go, A. S., Feldman, H. I., Mehta, R. C., et al. Association of tubular solute clearances with the glomerular filtration rate and complications of chronic kidney disease: the chronic renal insufficiency cohort study. *Nephrology Dialysis Transplantation 36*, 7 (2021), 1271–1281.

[28] Chen, Z., Li, J., Sun, Y., Wang, C., Yang, W., Ma, M., Luo, Z., Yang, K., and Chen, L. A novel predictive model for poor in-hospital outcomes in patients with acute kidney injury after cardiac surgery. *The Journal of Thoracic and Cardiovascular Surgery* (2021).

[29] Cleveland, W. S., and Devlin, S. J. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association 83*, 403 (1988), 596–610.

[30] Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. *Circulation 131*, 2 (2015), 211–219.

[31] Collobert, R., and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (2008), pp. 160–167.

[32] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[33] Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology 106*, 1 (2017), 1–9.

[34] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*, 2 (1972), 187–202.

[35] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*, 2 (1972), 187–220.

[36] Cox, D. R. Regression models and life-tables. breakthroughs in statistics, 1992.

[37] Cox, D. R., and Oakes, D. *Analysis of survival data*. Chapman and Hall/CRC, 2018.

[38] Dauda, K. A., Pradhan, B., Uma Shankar, B., and Mitra, S. Decision tree for modeling survival data with competing risks. *Biocybernetics and Biomedical Engineering 39*, 3 (2019), 697–708.

[39] Delanaye, P., Cavalier, E., Morel, J., Mehdi, M., Maillard, N., Claisse, G., Lambermont, B., Dubois, B. E., Damas, P., Krzesinski, J.-M., et al. Detection of decreased glomerular filtration rate in intensive care units: Serum cystatin c versusserum creatinine. *BMC nephrology 15*, 1 (2014), 1–6.

[40] Demirjian, S., Chertow, G. M., Zhang, J. H., O'Connor, T. Z., Vitale, J., Paganini, E. P., Palevsky, P. M., Network, V. A. R. F. T., et al. Model to predict mortality in critically ill adults with acute kidney injury. *Clinical Journal of the American Society of Nephrology 6*, 9 (2011), 2114–2120.

[41] Demšar, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research 7* (2006), 1–30.

[42] Deo, R. C. Machine learning in medicine. *Circulation 132*, 20 (2015), 1920–1930.

[43] Disease, K. Improving global outcomes (kdigo) acute kidney injury work group: Kdigo clinical practice guideline for acute kidney injury. *Kidney Int Suppl 2*, 1 (2012), 1–138.

[44] Dunn, O. J., and Clark, V. A. *Basic statistics: a primer for the biomedical sciences*. John Wiley & Sons, 2009.

[45] Fan, L., Levey, A. S., Gudnason, V., Eiriksdottir, G., Andresdottir, M. B., Gudmundsdottir, H., Indridason, O. S., Palsson, R., Mitchell, G., and Inker, L. A. Comparing gfr estimating equations using cystatin c and creatinine in elderly individuals. *Journal of the American Society of Nephrology 26*, 8 (2015), 1982–1989.

[46] Faraggi, D., and Simon, R. A neural network model for survival data. *Statistics in medicine 14*, 1 (1995), 73–82.

[47] Ferri, C., Hernández-Orallo, J., and Modroiu, R. An experimental comparison of performance measures for classification. *Pattern recognition letters 30*, 1 (2009), 27–38.

[48] Flechet, M., Güiza, F., Schetz, M., Wouters, P., Vanhorebeek, I., Derese, I., Gunst, J., Spriet, I., Casaer, M., Van den Berghe, G., et al. Akipredictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. *Intensive care medicine 43*, 6 (2017), 764–773.

[49] for Health Statistics, N. C. webpage. `https://wwwn.cdc.gov/nchs/nhanes/nhanes1/`.

[50] Forman, G., and Scholz, M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter 12*, 1 (2010), 49–57.

[51] Forni, L., Darmon, M., Ostermann, M., Oudemans-van Straaten, H., Pettilä, V., Prowle, J., Schetz, M., and Joannidis, M. Renal recovery after acute kidney injury. *Intensive care medicine 43*, 6 (2017), 855–866.

[52] Fuchs, L., Lee, J., Novack, V., Baumfeld, Y., Scott, D., Celi, L., Mandelbaum, T., Howell, M., and Talmor, D. Severity of acute kidney injury and two-year outcomes in critically ill patients. *Chest 144*, 3 (2013), 866–875.

[53] Fushiki, T. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing 21*, 2 (2011), 137–146.

[54] Geri, G., Stengel, B., Jacquelinet, C., Aegerter, P., Massy, Z. A., and Vieillard-Baron, A. Prediction of chronic kidney disease after acute kidney injury in icu patients: study protocol for the predict multicenter prospective observational study. *Annals of intensive care 8*, 1 (2018), 1–5.

[55] Gharaibeh, K. A., Hamadah, A. M., El-Zoghby, Z. M., Lieske, J. C., Larson, T. S., and Leung, N. Cystatin c predicts renal recovery earlier than creatinine among patients with acute kidney injury. *Kidney international reports 3*, 2 (2018), 337–342.

[56] Gordon, L., and Olshen, R. A. Tree-structured survival analysis. *Cancer treatment reports 69*, 10 (1985), 1065–1069.

[57] Grams, M. E., Waikar, S. S., MacMahon, B., Whelton, S., Ballew, S. H., and Coresh, J. Performance and limitations of administrative data in the identification of aki. *Clinical Journal of the American Society of Nephrology 9*, 4 (2014), 682–689.

[58] Grubb, A., Horio, M., Hansson, L.-O., Björk, J., Nyman, U., Flodin, M., Larsson, A., Bökenkamp, A., Yasuda, Y., Blufpand, H., et al. Generation of a new cystatin c–based estimating equation for glomerular filtration rate by use of 7 assays standardized to the international calibrator. *Clinical chemistry 60*, 7 (2014), 974–986.

[59] Gunter, T. D., and Terry, N. P. The emergence of national electronic health record architectures in the united states and australia: models, costs, and questions. *Journal of medical Internet research 7*, 1 (2005), e383.

[60] Hanley, J. A., and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*, 1 (1982), 29–36.

[61] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the yield of medical tests. *Jama 247*, 18 (1982), 2543–2546.

[62] Hashemian, S. M., Jamaati, H., Bidgoli, B. F., Farrokhi, F. R., Malekmohammad, M., Roozdar, S., Mohajerani, S. A., Bagheri, A., Radmnand, G., Hatami, B., et al. Outcome of acute kidney injury in critical care unit, based on aki network. *Tanaffos 15*, 2 (2016), 89.

[63] Hassanzadeh, H. R., Phan, J. H., and Wang, M. D. A multi-modal graph-based semi-supervised pipeline for predicting cancer survival. In

2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2016), IEEE, pp. 184–189.

[64] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[65] Hayden, J. A., Côté, P., and Bombardier, C. Evaluation of the quality of prognosis studies in systematic reviews. *Annals of internal medicine 144*, 6 (2006), 427–437.

[66] He, J., Lin, J., and Duan, M. Application of machine learning to predict acute kidney disease in patients with sepsis associated acute kidney injury. *Frontiers in Medicine 8* (2021).

[67] Helmersson-Karlqvist, J., Lipcsey, M., Ärnlöv, J., Bell, M., Ravn, B., Dardashti, A., and Larsson, A. Cystatin c predicts long term mortality better than creatinine in a nationwide study of intensive care patients. *Scientific Reports 11*, 1 (2021), 1–9.

[68] Hilden, J., Habbema, J. D. F., and Bjerregaard, B. The measurement of performance in probabilistic diagnosis. *Methods of information in medicine 17*, 04 (1978), 227–237.

[69] Hobson, C. E., Yavas, S., Segal, M. S., Schold, J. D., Tribble, C. G., Layon, A. J., and Bihorac, A. Acute kidney injury is associated with increased long-term mortality after cardiothoracic surgery. *Circulation 119*, 18 (2009), 2444–2453.

[70] Hoerl, A. E., and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1970), 55–67.

[71] Hosmer, D. W. Assessing the fit of the model. *Applied logistic regression* (2000), 143–202.

[72] Hosmer Jr, D. W., Lemeshow, S., and May, S. *Applied survival analysis: regression modeling of time-to-event data*, vol. 618. John Wiley & Sons, 2011.

[73] Hoste, E. A., Bagshaw, S. M., Bellomo, R., Cely, C. M., Colman, R., Cruz, D. N., Edipidis, K., Forni, L. G., Gomersall, C. D., Govil, D., et al. Epidemiology of acute kidney injury in critically ill patients: the multinational aki-epi study. *Intensive care medicine 41*, 8 (2015), 1411–1423.

[74] Hoste, E. A., Clermont, G., Kersten, A., Venkataraman, R., Angus, D. C., De Bacquer, D., and Kellum, J. A. Rifle criteria for

acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis. *Critical care 10*, 3 (2006), R73.

[75] Hoste, E. A., Kellum, J. A., Selby, N. M., Zarbock, A., Palevsky, P. M., Bagshaw, S. M., Goldstein, S. L., Cerdá, J., and Chawla, L. S. Global epidemiology and outcomes of acute kidney injury. *Nature Reviews Nephrology 14*, 10 (2018), 607–625.

[76] Hoste, L., Dubourg, L., Selistre, L., De Souza, V. C., Ranchin, B., Hadj-Aïssa, A., Cochat, P., Martens, F., and Pottel, H. A new equation to estimate the glomerular filtration rate in children, adolescents and young adults. *Nephrology Dialysis Transplantation 29*, 5 (2014), 1082–1091.

[77] Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. Bagging survival trees. *Statistics in medicine 23*, 1 (2004), 77–91.

[78] Houthooft, R., Ruyssinck, J., van der Herten, J., Stijven, S., Couckuyt, I., Gadeyne, B., Ongenae, F., Colpaert, K., Decruyenaere, J., Dhaene, T., et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. *Artificial intelligence in medicine 63*, 3 (2015), 191–207.

[79] Hsu, C.-y., Yang, W., Parikh, R. V., Anderson, A. H., Chen, T. K., Cohen, D. L., He, J., Mohanty, M. J., Lash, J. P., Mills, K. T., et al. Race, genetic ancestry, and estimating kidney function in ckd. *New England Journal of Medicine 385*, 19 (2021), 1750–1760.

[80] Hsu, R. K., and Hsu, C.-y. The role of acute kidney injury in chronic kidney disease. In *Seminars in nephrology* (2016), vol. 36, Elsevier, pp. 283–292.

[81] Hu, S., Fridgeirsson, E., van Wingen, G., and Welling, M. Transformer-based deep survival analysis. In *Survival Prediction-Algorithms, Challenges and Applications* (2021), PMLR, pp. 132–148.

[82] Huang, C.-Y., Güiza, F., De Vlieger, G., Wouters, P., Gunst, J., Casaer, M., Vanhorebeek, I., Derese, I., Van den Berghe, G., and Meyfroidt, G. Development and validation of clinical prediction models for acute kidney injury recovery at hospital discharge in critically ill adults. *Journal of Clinical Monitoring and Computing* (2022), 1–13.

[83] Inker, L. A., Eneanya, N. D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., Crews, D. C., Doria, A., Estrella, M. M., Froissart, M., et al. New creatinine-and cystatin c–based equations to estimate gfr without race. *New England Journal of Medicine 385*, 19 (2021), 1737–1749.

[84] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. Random survival forests. *The annals of applied statistics 2*, 3 (2008), 841–860.

[85] Itenov, T. S., Berthelsen, R. E., Jensen, J.-U., Gerds, T. A., Pedersen, L. M., Strange, D., Thormar, K., Løken, J., Andersen, M. H., Tousi, H., et al. Predicting recovery from acute kidney injury in critically ill patients: development and validation of a prediction model. *Critical Care and Resuscitation 20*, 1 (2018), 54.

[86] James, M. T., Pannu, N., Hemmelgarn, B. R., Austin, P. C., Tan, Z., McArthur, E., Manns, B. J., Tonelli, M., Wald, R., Quinn, R. R., et al. Derivation and external validation of prediction models for advanced chronic kidney disease following acute kidney injury. *Jama 318*, 18 (2017), 1787–1797.

[87] Järvisalo, M. J., Kartiosuo, N., Hellman, T., and Uusalo, P. Predicting mortality in critically ill patients requiring renal replacement therapy for acute kidney injury in a retrospective single-center study of two cohorts. *Scientific Reports 12*, 1 (Jun 2022), 10177.

[88] Kalra, D. Electronic health record standards. *Yearbook of medical informatics 15*, 01 (2006), 136–144.

[89] Kaplan, E. L., and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*, 282 (1958), 457–481.

[90] Kar, S., Paglialunga, S., and Islam, R. Cystatin c is a more reliable biomarker for determining egfr to support drug development studies. *The Journal of Clinical Pharmacology 58*, 10 (2018), 1239–1247.

[91] Kawaler, E., Cobian, A., Peissig, P., Cross, D., Yale, S., and Craven, M. Learning to predict post-hospitalization vte risk from ehr data. In *AMIA annual symposium proceedings* (2012), vol. 2012, American Medical Informatics Association, p. 436.

[92] Kellum, J. A., Lameire, N., Aspelin, P., Barsoum, R. S., Burdmann, E. A., Goldstein, S. L., Herzog, C. A., Joannidis, M., Kribben, A., Levey, A. S., et al. Kidney disease: improving global outcomes (kdigo) acute kidney injury work group. kdigo clinical practice guideline for acute kidney injury. *Kidney international supplements 2*, 1 (2012), 1–138.

[93] Kellum, J. A., and Prowle, J. R. Paradigms of acute kidney injury in the intensive care setting. *Nature Reviews Nephrology 14*, 4 (2018), 217.

[94] Kennedy, E. H., Wiitala, W. L., Hayward, R. A., and Sussman, J. B. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical care 51*, 3 (2013), 251.

[95] Kerr, M., Bedford, M., Matthews, B., and O'Donoghue, D. The economic impact of acute kidney injury in england. *Nephrology Dialysis Transplantation 29*, 7 (2014), 1362–1368.

[96] Khan, F. M., and Zubek, V. B. Support vector regression for censored data (svrc): a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining* (2008), IEEE, pp. 863–868.

[97] Khwaja, A. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice 120*, 4 (2012), c179–c184.

[98] Kimble, C. Electronic health records: Cure-all or chronic condition? *Global Business and Organizational Excellence 33*, 4 (2014), 63–74.

[99] Kleinbaum, D. G., and Klein, M. *Survival analysis*. Springer, 2010.

[100] Kleinbaum, D. G., Klein, M., et al. *Survival analysis: a self-learning text*, vol. 3. Springer, 2012.

[101] Knight, E. L., Verhave, J. C., Spiegelman, D., Hillege, H. L., De Zeeuw, D., Curhan, G. C., and De Jong, P. E. Factors influencing serum cystatin c levels other than renal function and the impact on renal function measurement. *Kidney international 65*, 4 (2004), 1416–1421.

[102] Koyner, J. L., Carey, K. A., Edelson, D. P., and Churpek, M. M. The development of a machine learning inpatient acute kidney injury prediction model. *Critical care medicine 46*, 7 (2018), 1070–1077.

[103] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature 521*, 7553 (2015), 436–444.

[104] Lee, B. J., Hsu, C.-y., Parikh, R., McCulloch, C. E., Tan, T. C., Liu, K. D., Hsu, R. K., Pravoverov, L., Zheng, S., and Go, A. S. Predicting renal recovery after dialysis-requiring acute kidney injury. *Kidney international reports 4*, 4 (2019), 571–581.

[105] Lee, E. T., and Wang, J. *Statistical methods for survival data analysis*, vol. 476. John Wiley & Sons, 2003.

[106] Lee, H.-C., Yoon, H.-K., Nam, K., Cho, Y. J., Kim, T. K., Kim, W. H., and Bahk, J.-H. Derivation and validation of machine learning

approaches to predict acute kidney injury after cardiac surgery. *Journal of clinical medicine 7*, 10 (2018), 322.

[107] Levatić, J., Ceci, M., Kocev, D., and Džeroski, S. Semi-supervised classification trees. *Journal of Intelligent Information Systems 49*, 3 (2017), 461–486.

[108] Levey, A., Atkins, R., Coresh, J., Cohen, E., Collins, A., Eckardt, K.-U., Nahas, M., Jaber, B., Jadoul, M., Levin, A., et al. Chronic kidney disease as a global public health problem: approaches and initiatives–a position statement from kidney disease improving global outcomes. *Kidney international 72*, 3 (2007), 247–259.

[109] Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N., Roth, D., and of Diet in Renal Disease Study Group*, M. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of internal medicine 130*, 6 (1999), 461–470.

[110] Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y., Castro III, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T., et al. A new equation to estimate glomerular filtration rate. *Annals of internal medicine 150*, 9 (2009), 604–612.

[111] Levin, A., Stevens, P. E., Bilous, R. W., Coresh, J., De Francisco, A. L., De Jong, P. E., Griffith, K. E., Hemmelgarn, B. R., Iseki, K., Lamb, E. J., et al. Kidney disease: Improving global outcomes (kdigo) ckd work group. kdigo 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international supplements 3*, 1 (2013), 1–150.

[112] Li, D. H., Wald, R., Blum, D., McArthur, E., James, M. T., Burns, K. E., Friedrich, J. O., Adhikari, N. K., Nash, D. M., Lebovic, G., Harvey, A. K., Dixon, S. N., Silver, S. A., Bagshaw, S. M., and Beaubien-Souligny, W. Predicting mortality among critically ill patients with acute kidney injury treated with renal replacement therapy: Development and validation of new prediction models. *Journal of Critical Care 56* (2020), 113–119.

[113] Li, M., and Zhou, Z.-H. Setred: Self-training with editing. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2005), Springer, pp. 611–621.

[114] Li, Y., Wang, L., Wang, J., Ye, J., and Reddy, C. K. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox

regression. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (2016), IEEE, pp. 231–240.

[115] Liang, Y., Chai, H., Liu, X.-Y., Xu, Z.-B., Zhang, H., and Leung, K.-S. Cancer survival analysis using semi-supervised learning method based on cox and aft models with l 1/2 regularization. *BMC medical genomics 9*, 1 (2016), 1–11.

[116] Liangos, O., Wald, R., O'Bell, J. W., Price, L., Pereira, B. J., and Jaber, B. L. Epidemiology and outcomes of acute renal failure in hospitalized patients: a national survey. *Clinical journal of the American Society of Nephrology 1*, 1 (2006), 43–51.

[117] Luo, G. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics 5*, 1 (2016), 1–16.

[118] Madani, A., Moradi, M., Karargyris, A., and Syeda-Mahmood, T. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)* (2018), IEEE, pp. 1038–1042.

[119] Makris, K. The role of the clinical laboratory in the detection and monitoring of acute kidney injury. *J Lab Precis Med 3* (2018), 69.

[120] Makris, K., and Spanou, L. Acute kidney injury: definition, pathophysiology and clinical phenotypes. *The Clinical Biochemist Reviews 37*, 2 (2016), 85.

[121] Makris, K., and Spanou, L. Acute kidney injury: diagnostic approaches and controversies. *The Clinical Biochemist Reviews 37*, 4 (2016), 153.

[122] McClosky, D., Charniak, E., and Johnson, M. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (2006), pp. 152–159.

[123] McDonald, H. I., Shaw, C., Thomas, S. L., Mansfield, K. E., Tomlinson, L. A., and Nitsch, D. Methodological challenges when carrying out research on ckd and aki using routine electronic health records. *Kidney international 90*, 5 (2016), 943–949.

[124] Mehta, R. L., Kellum, J. A., Shah, S. V., Molitoris, B. A., Ronco, C., Warnock, D. G., and Levin, A. Acute kidney injury network: report of an initiative to improve outcomes in acute kidney injury. *Critical care 11*, 2 (2007), 1–8.

[125] Miao, F., Cai, Y.-P., Zhang, Y.-T., and Li, C.-Y. Is random survival forest an alternative to cox proportional model on predicting cardiovascular disease? In *6TH European conference of the international federation for medical and biological engineering* (2015), Springer, pp. 740–743.

[126] Miller, R. G. *Survival Analysis*. Wiley-Blackwell, 1981.

[127] Mitchell, T. M., and Mitchell, T. M. *Machine learning*, vol. 1. McGraw-hill New York, 1997.

[128] Mohamadlou, H., Lynn-Palevsky, A., Barton, C., Chettipally, U., Shieh, L., Calvert, J., Saber, N. R., and Das, R. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Canadian journal of kidney health and disease 5* (2018), 2054358118776326.

[129] Mohsenin, V. Practical approach to detection and management of acute kidney injury in critically ill patient. *Journal of intensive care 5*, 1 (2017), 1–8.

[130] Molinaro, A. M., Simon, R., and Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics 21*, 15 (2005), 3301–3307.

[131] Murphy, K. P. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[132] Murugan, R., Karajala-Subramanyam, V., Lee, M., Yende, S., Kong, L., Carter, M., Angus, D. C., Kellum, J. A., et al. Acute kidney injury in non-severe pneumonia is associated with an increased immune response and lower survival. *Kidney international 77*, 6 (2010), 527–535.

[133] Nateghi Haredasht, F., Antonatou, M., Cavalier, E., Delanaye, P., Pottel, H., and Makris, K. The effect of different consensus definitions on diagnosing acute kidney injury events and their association with in-hospital mortality. *Journal of Nephrology* (2022), 1–9.

[134] Nateghi Haredasht, F., and Vens, C. Predicting survival outcomes in the presence of unlabeled data. *Machine Learning* (2022), 1–19.

[135] Nateghi Haredasht, F., Viaene, L., De Corte, W., and Vens, C. Exploiting unlabeled data to predict the development of ckd after aki in critically ill patients. In *Intelligence Artificielle & Néphrologie 2021, Date: 2021/02/11-2021/02/12, Location: Paris, France* (2021).

[136] Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., and Buchman, T. G. An interpretable machine learning model

for accurate prediction of sepsis in the icu. *Critical care medicine 46*, 4 (2018), 547.

[137] Network, V. A. R. F. T. Intensity of renal support in critically ill patients with acute kidney injury. *New England Journal of Medicine 359*, 1 (2008), 7–20.

[138] Nie, S., Tang, L., Zhang, W., Feng, Z., and Chen, X. Are there modifiable risk factors to improve aki? *BioMed research international 2017* (2017).

[139] Nigam, K., and Ghani, R. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management* (2000), pp. 86–93.

[140] Nisula, S., Kaukonen, K.-M., Vaara, S. T., Korhonen, A.-M., Poukkanen, M., Karlsson, S., Haapio, M., Inkinen, O., Parviainen, I., Suojaranta-Ylinen, R., et al. Incidence, risk factors and 90-day mortality of patients with acute kidney injury in finnish intensive care units: the finnaki study. *Intensive care medicine 39*, 3 (2013), 420–428.

[141] Opitz, D., and Maclin, R. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research 11* (1999), 169–198.

[142] Ostermann, M., and Cerdá, J. The burden of acute kidney injury and related financial issues. *Acute Kidney Injury-Basic Research and Clinical Practice 193* (2018), 100–112.

[143] Palevsky, P. M., Zhang, J. H., O'Connor, T. Z., Chertow, G. M., Crowley, S. T., Choudhury, D., Finkel, K., Kellum, J. A., Paganini, E., Schein, R. M., et al. Intensity of renal support in critically ill patients with acute kidney injury (new england journal of medicine (2008) 359,(7-20)). *New England Journal of Medicine 361*, 24 (2009).

[144] Panahiazar, M., Taslimitehrani, V., Pereira, N., and Pathak, J. Using ehrs and machine learning for heart failure survival analysis. *Studies in health technology and informatics 216* (2015), 40.

[145] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research 12* (2011), 2825–2830.

[146] Petrucelli, N., Daly, M. B., and Pal, T. Brca1-and brca2-associated hereditary breast and ovarian cancer.

[147] Pike, F., Murugan, R., Keener, C., Palevsky, P. M., Vijayan, A., Unruh, M., Finkel, K., Wen, X., and Kellum, J. A. Biomarker enhanced risk prediction for adverse outcomes in critically ill patients receiving rrt. *Clinical Journal of the American Society of Nephrology 10*, 8 (2015), 1332–1339.

[148] Pottel, H., Björk, J., Courbebaisse, M., Couzi, L., Ebert, N., Eriksen, B. O., Dalton, R. N., Dubourg, L., Gaillard, F., Garrouste, C., et al. Development and validation of a modified full age spectrum creatinine-based equation to estimate glomerular filtration rate: a cross-sectional analysis of pooled data. *Annals of internal medicine 174*, 2 (2021), 183–191.

[149] Pottel, H., Delanaye, P., Schaeffner, E., Dubourg, L., Eriksen, B. O., Melsom, T., Lamb, E. J., Rule, A. D., Turner, S. T., Glassock, R. J., et al. Estimating glomerular filtration rate for the full age spectrum from serum creatinine and cystatin c. *Nephrology Dialysis Transplantation 32*, 3 (2017), 497–507.

[150] Pottel, H., Hoste, L., Dubourg, L., Ebert, N., Schaeffner, E., Eriksen, B. O., Melsom, T., Lamb, E. J., Rule, A. D., Turner, S. T., et al. An estimated glomerular filtration rate equation for the full age spectrum. *Nephrology Dialysis Transplantation 31*, 5 (2016), 798–806.

[151] Price, C. P., and Finney, H. Developments in the assessment of glomerular filtration rate. *Clinica chimica acta 297*, 1-2 (2000), 55–66.

[152] Prowle, J. R., Kolic, I., Purdell-Lewis, J., Taylor, R., Pearse, R. M., and Kirwan, C. J. Serum creatinine changes associated with critical illness and detection of persistent renal dysfunction after aki. *Clinical Journal of the American Society of Nephrology 9*, 6 (2014), 1015–1023.

[153] Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., and Fanos, V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina 56*, 9 (2020), 455.

[154] Rayan, Z., Alfonse, M., and Salem, A.-B. M. Intensive care unit (icu) data analytics using machine learning techniques. *Int J Inf Theor Appl 26*, 1 (2019), 69–82.

[155] Rewa, O., and Bagshaw, S. M. Acute kidney injury—epidemiology, outcomes and economics. *Nature reviews nephrology 10*, 4 (2014), 193–207.

[156] Rimes-Stigare, C., Ravn, B., Awad, A., Torlén, K., Martling, C.-R., Bottai, M., Mårtensson, J., and Bell, M. Creatinine-and

cystatin c-based incidence of chronic kidney disease and acute kidney disease in aki survivors. *Critical care research and practice 2018* (2018).

[157] Rogers, T., Worden, K., Fuentes, R., Dervilis, N., Tygesen, U., and Cross, E. A bayesian non-parametric clustering approach for semi-supervised structural health monitoring. *Mechanical Systems and Signal Processing 119* (2019), 100–119.

[158] Roli, F., and Marcialis, G. L. Semi-supervised pca-based face recognition using self-training. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (2006), Springer, pp. 560–568.

[159] Ronco, C., Bellomo, R., and Kellum, J. A. Acute kidney injury. *The Lancet 394*, 10212 (2019), 1949–1964.

[160] Roy, B., Stepišnik, T., ALS, T. P. R. O.-A., Vens, C., Džeroski, S., Consortium, C. T., et al. Survival analysis with semi-supervised predictive clustering trees. *Computers in Biology and Medicine 141* (2022), 105001.

[161] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*, 5 (2019), 206–215.

[162] Sakr, S., Elshawi, R., Ahmed, A., Qureshi, W. T., Brawner, C., Keteyian, S., Blaha, M. J., and Al-Mallah, M. H. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The henry ford exercise testing (fit) project. *PLoS One 13*, 4 (2018), e0195344.

[163] Sanchez-Pinto, L. N., Luo, Y., and Churpek, M. M. Big data and data science in critical care. *Chest 154*, 5 (2018), 1239–1248.

[164] Sato, Y., Takahashi, M., and Yanagita, M. Pathophysiology of aki to ckd progression. In *Seminars in Nephrology* (2020), vol. 40, Elsevier, pp. 206–215.

[165] Schmid, M., Wright, M. N., and Ziegler, A. On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications 63* (2016), 450–459.

[166] Segal, M. R. Regression trees for censored data. *Biometrics* (1988), 35–47.

[167] Shailaja, K., Seetharamulu, B., and Jabbar, M. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)* (2018), IEEE, pp. 910–914.

[168] Shemesh, O., Golbetz, H., Kriss, J. P., and Myers, B. D. Limitations of creatinine as a filtration marker in glomerulopathic patients. *Kidney international 28*, 5 (1985), 830–838.

[169] Shi, M., and Zhang, B. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics 27*, 21 (2011), 3017–3023.

[170] Shillan, D., Sterne, J. A., Champneys, A., and Gibbison, B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical care 23*, 1 (2019), 1–11.

[171] Siew, E. D., and Davenport, A. The growth of acute kidney injury: a rising tide or just closer attention to detail? *Kidney international 87*, 1 (2015), 46–61.

[172] Sileanu, F. E., Murugan, R., Lucko, N., Clermont, G., Kane-Gill, S. L., Handler, S. M., and Kellum, J. A. Aki in low-risk versus high-risk patients in intensive care. *Clinical Journal of the American Society of Nephrology 10*, 2 (2015), 187–196.

[173] Silver, S. A., and Chertow, G. M. The economic consequences of acute kidney injury. *Nephron 137*, 4 (2017), 297–301.

[174] Silver, S. A., Long, J., Zheng, Y., and Chertow, G. M. Cost of acute kidney injury in hospitalized patients. *Journal of hospital medicine 12*, 2 (2017), 70–76.

[175] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software 39*, 5 (2011), 1.

[176] Sohaney, R., Yin, H., Shahinian, V., Saran, R., Steffick, D., Nallamothu, B. K., and Heung, M. Trends in the incidence of acute kidney injury in a national cohort of us veterans. *American Journal of Kidney Diseases 77*, 2 (2021), 300–302.

[177] Sparrow, H. G., Swan, J. T., Moore, L. W., Gaber, A. O., and Suki, W. N. Disparate outcomes observed within kidney disease: Improving global outcomes (kdigo) acute kidney injury stage 1. *Kidney international 95*, 4 (2019), 905–913.

[178] Srisawat, N., Wen, X., Lee, M., Kong, L., Elder, M., Carter, M., Unruh, M., Finkel, K., Vijayan, A., Ramkumar, M., et al. Urinary biomarkers and renal recovery in critically ill patients with renal support. *Clinical Journal of the American Society of Nephrology 6*, 8 (2011), 1815–1823.

[179] Stack, A. G., Li, X., Kaballo, M. A., Elsayed, M. E., Johnson, H., Murray, P. T., Saran, R., and Browne, L. D. Temporal trends in acute kidney injury across health care settings in the irish health system: a cohort study. *Nephrology Dialysis Transplantation 35*, 3 (2020), 447–457.

[180] Steck, H., Krishnapuram, B., Dehing-Oberije, C., Lambin, P., and Raykar, V. C. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems* (2008), pp. 1209–1216.

[181] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj 338* (2009).

[182] Stevens, L. A., Zhang, Y. L., and Schmid, C. H. Evaluating the performance of gfr estimating equations. *Journal of nephrology 21*, 6 (2008), 797.

[183] Survlab. webpage. `http://user.it.uu.se/~kripe367/survlab/download.html`, 2010 (accessed December 7, 2014).

[184] Tanha, J., van Someren, M., and Afsarmanesh, H. Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics 8*, 1 (2017), 355–370.

[185] Therneau, T. M. *A Package for Survival Analysis in R*, 2020. R package version 3.2-7.

[186] Thomas, M. E., Blaine, C., Dawnay, A., Devonald, M. A., Ftouh, S., Laing, C., Latchem, S., Lewington, A., Milford, D. V., and Ostermann, M. The definition of acute kidney injury and its use in practice. *Kidney international 87*, 1 (2015), 62–73.

[187] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[188] Tibshirani, R. The lasso method for variable selection in the cox model. *Statistics in medicine 16*, 4 (1997), 385–395.

[189] Tidman, M., Sjöström, P., and Jones, I. A comparison of gfr estimating formulae based upon s-cystatin c and s-creatinine and a combination of the two. *Nephrology Dialysis Transplantation 23*, 1 (2008), 154–160.

[190] Triguero, I., García, S., and Herrera, F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems 42*, 2 (2015), 245–284.

[191] Uchino, S., Kellum, J. A., Bellomo, R., Doig, G. S., Morimatsu, H., Morgera, S., Schetz, M., Tan, I., Bouman, C., Macedo, E., et al. Acute renal failure in critically ill patients: a multinational, multicenter study. *Jama 294*, 7 (2005), 813–818.

[192] Vaidya, V. S., Ferguson, M. A., and Bonventre, J. V. Biomarkers of acute kidney injury. *Annual review of pharmacology and toxicology 48* (2008), 463.

[193] Van Buuren, S., and Oudshoorn, C. G. Multivariate imputation by chained equations, 2000.

[194] van Engelen, J. E., and Hoos, H. H. A survey on semi-supervised learning. *Machine Learning* (Nov 2019).

[195] Van Engelen, J. E., and Hoos, H. H. A survey on semi-supervised learning. *Machine Learning 109*, 2 (2020), 373–440.

[196] Vanneschi, L., Farinaccio, A., Mauri, G., Antoniotti, M., Provero, P., and Giacobini, M. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData mining 4*, 1 (2011), 1–13.

[197] Vickers, A. J., and Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making 26*, 6 (2006), 565–574.

[198] Vinzamuri, B., Li, Y., and Reddy, C. K. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (2014), pp. 241–250.

[199] Vranas, K. C., Jopling, J. K., Sweeney, T. E., Ramsey, M. C., Milstein, A. S., Slatore, C. G., Escobar, G. J., and Liu, V. X. Identifying distinct subgroups of intensive care unit patients: A machine learning approach. *Critical care medicine 45*, 10 (2017), 1607.

[200] Wald, R., Quinn, R. R., Luo, J., Li, P., Scales, D. C., Mamdani, M. M., Ray, J. G., of Toronto Acute Kidney Injury Research Group, U., et al. Chronic dialysis and death among survivors of acute kidney injury requiring dialysis. *Jama 302*, 11 (2009), 1179–1185.

[201] Wang, H. E., Muntner, P., Chertow, G. M., and Warnock, D. G. Acute kidney injury and mortality in hospitalized patients. *American journal of nephrology 35*, 4 (2012), 349–355.

[202] Wang, P., Li, Y., and Reddy, C. K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR) 51*, 6 (2019), 1–36.

[203] Wang, Z., and Sun, J. Survtrace: transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (2022), pp. 1–9.

[204] Warnock, D. G., Neyra, J. A., Macedo, E., Miles, A. D., Mehta, R. L., and Wanner, C. Comparison of static and dynamic baseline creatinine surrogates for defining acute kidney injury. *Nephron 145*, 6 (2021), 664–674.

[205] Wei, C., Zhang, L., Feng, Y., Ma, A., and Kang, Y. Machine learning model for predicting acute kidney injury progression in critically ill patients. *BMC medical informatics and decision making 22*, 1 (2022), 1–11.

[206] Werner, K., Pihlsgård, M., Elmståhl, S., Legrand, H., Nyman, U., and Christensson, A. Combining cystatin c and creatinine yields a reliable glomerular filtration rate estimation in older adults in contrast to β-trace protein and β2-microglobulin. *Nephron 137*, 1 (2017), 29–37.

[207] Wilson, T., Quan, S., Cheema, K., Zarnke, K., Quinn, R., de Koning, L., Dixon, E., Pannu, N., and James, M. T. Risk prediction models for acute kidney injury following major noncardiac surgery: systematic review. *Nephrology Dialysis Transplantation 31*, 2 (2016), 231–240.

[208] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., and DATA, M. Practical machine learning tools and techniques. In *Data Mining* (2005), vol. 2.

[209] Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics* (1995), pp. 189–196.

[210] Yu, K.-H., Beam, A. L., and Kohane, I. S. Artificial intelligence in healthcare. *Nature biomedical engineering 2*, 10 (2018), 719–731.

[211] Závada, J., Hoste, E., Cartin-Ceba, R., Calzavacca, P., Gajic, O., Clermont, G., Bellomo, R., Kellum, J. A., and investigators, A. A comparison of three methods to estimate baseline

creatinine for rifle classification. *Nephrology Dialysis Transplantation 25*, 12 (2010), 3911–3918.

[212] Zeng, X., McMahon, G. M., Brunelli, S. M., Bates, D. W., and Waikar, S. S. Incidence, outcomes, and comparisons across definitions of aki in hospitalized individuals. *Clinical Journal of the American Society of Nephrology 9*, 1 (2014), 12–20.

[213] Zhu, X. J. Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 2005.

[214] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*, 2 (2005), 301–320.

[215] Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R., and Bratko, I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine 20*, 1 (2000), 59–75.

[216] Zupan, B., Demšar, J., Kattan, M. W., Beck, J., and Bratko, I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine 20*, 1 (2000), 59–75. Selected Papers from AIMDM '99.

Curriculum vitae

Fateme Nateghi Haredasht was born in Iran on June 28, 1988. She received her bachelor's degree in Electrical Engineering from the Univeridty of Guilan in 2012. Further, she attended a master's program in communications engineering; however, she decided to quit the program since she acquired a strong interest in machine learning and its applications in the medical area. Afterward, she continued her academic training with a master's degree in Medical Informatics at the Amirkabir University of Technology in 2016. After two years of doing research in the field of machine learning, in a collaborative paper with her colleagues, she managed to publish a journal paper in Pattern Recognition journal.

In January 2019, she joined KU Leuven, Faculty of Medicine at Campus KULAK to pursue a PhD in machine learning in healthcare applications, under the supervision of prof. Celine Vens, prof. Hans Pottel, dr. Wouter De Corte, and dr. Liesbeth Viaene. Her research interests lie in the area of machine learning with applications in healthcare. More specifically, she focuses on adapting semi-supervised learning models to the domain of survival analysis. Her project aimed to develop novel machine learning-based prediction models to be used in the ICU setting with healthcare data. She was able to publish 2 scientific articles regarding her PhD work and submit 4 more that are currently under review. Also, a scientific article is currently in preparation as a result of her participation as a collaborator in projects.

List of publications

Articles in internationally reviewed academic journals

Ashtari P., **Nateghi Haredasht, F.**, and Beigy H., Supervised fuzzy partitioning. Pattern Recognition 2020, 97, p.107013.9.

Nateghi Haredasht, F., Antonatou M., CavalierE., Delanaye P., Pottel P., and Makris K., The effect of different consensus definitions on diagnosing acute kidney injury events and their association with in-hospital mortality. Journal of Nephrology 2022: 1-9.

Nateghi Haredasht, F. and Vens, C., Predicting Survival Outcomes in the Presence of Unlabeled Data. Machine Learning (2022): 1-19.

Nateghi Haredasht, F., Viaene L., Vens C., Callewaert N., De Corte W., and Pottel H., Comparison between cystatin C- and creatinine-based estimated glomerular filtration rate in the follow-up of patients recovering from a stage 3 AKI in ICU. Journal of Clinical Medicine 11, no. 24 (2022): 7264.

Nateghi Haredasht, F., Vanhoutte L., Vens C., Pottel H., Viaene L., and De Corte W., Validated risk prediction models for outcomes of acute kidney injury: a systematic review (submitted to BMC Nephrology).

Nateghi Haredasht, F., Dauda K.A., and Vens C., Exploiting censored information in self-training for time-to-event prediction (submitted to the International Journal of Medical Informatics).

Nateghi Haredasht, F., Viaene L., Pottel H., De Corte W., and Vens C., Predicting outcomes of acute kidney injury in critically ill patients using machine learning (submitted to the Journal of the American Medical Informatics Association : JAMIA).

Papers at scientific conferences, published in full

Nateghi Haredasht, F., Ghassemi F., and Moradi M.H., Causal inference of gene expression data using a clustering-based extension of Kernel-Granger causality. 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME). IEEE, 2016.

Abstracts, presented at scientific conferences

Nateghi Haredasht, F., and Moradi M.H., Nonlinear Causality Inference in Microarray Time Series. BNAIC/BENELEARN 2019.

Nateghi Haredasht, F., Viaene L., De Corte W., and Vens C., Exploiting unlabeled data to predict the development of CKD after AKI in critically ill patients. Intelligence Artificielle & Nephrologie, Paris 2021, France.

Nateghi Haredasht, F., Makris K., Delanaye P., and Pottel H., Association of AKI-event defined by Different Definitions with In-hospital Mortality. ERA-EDTA 2021.

Nateghi Haredasht, F., Liesbeth L., Vens, C., De Corte W., Pottel H., Serum Creatinine-based versus Cystatin C-based eGFR in AKI-stage3 critically ill patients. Annual Congress of the European Society of Intensive Care Medicine (ESICM) 2022, Madrid, Spain.

Scientific acknowledgement The scientific acknowledgments have been stated in the corresponding chapters (research articles) of this thesis.

Personal contribution This PhD thesis was written by Fateme Nateghi Haredasht. It was reviewed by the PhD supervisor Prof. Celine Vens and PhD co-supervisors Prof. Hans Pottel, Dr. Wouter De Corte, and Dr. Liesbeth Viaene.

All experiments were performed by the PhD candidate. All (co)-supervisors Prof. Celine Vens, Prof. Hans Pottel, Dr. Wouter De Corte, and Dr. Liesbeth Viaene provided valuable guidance throughout this PhD. The data used in this thesis are publicly available at[1], except the data related to Chapters 4, 6, and 9 where legal restrictions apply. The PhD candidate had a substantial contribution to the conception and design of the performed studies.

Conflict of interest

No conflict of interest is declared.

[1] https://fatemenateghi.github.io/

FACULTY OF MEDICINE
DEPARTMENT OF PUBLIC HEALTH AND PRIMARY CARE
BIOMEDICAL SCIENCES GROUP, ITEC-IMEC RESEARCH GROUP AT KU LEUVEN
Oude Markt 13
3000 Leuven
fateme.nateghi@kuleuven.be