MULTI-STATE CLINICAL PREDICTION MODELS IN RENAL REPLACEMENT THERAPY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2022

By
Michael A Barrowman
School of Health Sciences

Blank Page

Contents

Word Count: 38,249

List of Tables

List of Figures

Blank Page

Abstract

The aim of this thesis is to explore and quantify properties associated with Multi-State Clinical Prediction Modelling.

We use simulations to infer knowledge regarding causal assumptions in competing risks scenarios (a subset of Multi-State Models) and time-dependent measures of model calibration. The causal assessment involves the investigation of multiple real-world scenarios where confounding factors may interfere with the standard way of measuring the effects of a treatment on an event-of-interest and a competing event. This is important in the field of Multi-State Models as the inaccurate interpretation of an effect on a competing event can lead to misconceptions in the causes of the event-of-interest.

Further simulations are performed to analyse how traditional methods of assessing the Calibration-in-the-Large of a Clinical Prediction Model can be affected by the censoring of patients over time, in particular when this censoring is caused by a competing event related to the variable of interest. To combat this, we use the Inverse Probability of Censoring to weight patients based on their likelihood to still be present in the data at a certain time, to re-align the measurements with reality and avoid bias due any underlying relationship between the competing event and the attributes of a patient.

This knowledge feeds into the design and implementation of metrics to assess other aspects of model validity, namely accuracy, discrimination and calibration, in a Multi-State Clinical Prediction Model. The Brier Score is extended to account for multiple outcomes, and the c-statistic is replaced by the Polytomous Discriminatory Index. Both of these extended measures are adjusted to fit into the scales of their traditional counterparts. We also extend the measures of calibration (i.e. Intercept and Slope), and encode further information into these metrics by also analysing the traditionally held assumption of that state predictions are completely independent. All of these methods are augmented with the information garnered from the previous simulations to ensure that bias due to censoring is accounted for.

Data from the Salford Kidney Study and the West of Scotland Electronic Renal Patient Record are used to develop and validate our own clinical prediction model. This model can predict a Chronic Kidney Disease patient's journey through Renal Replacement Therapy and on to Death, and through the application of our validity metrics, we can be confident that it can be accurate and effective in its predictions.

Blank Page

Declaration

that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Blank Page

Copyright

Blank Page

Acknowledgements

My everlasting gratitude will always lie with my parents, Brenda and Andy, who have allowed me to be myself and to find my own path. They never pushed me into something that I didn't want to do, but also guided me towards what I would be good at. They gave me the freedom to learn who I am as a person and the solid support to achieve what I needed. As clichéd as it sounds, I would not be the person I am today if it were not for them.

My thanks also go to my children, Matilda and Theodore, who have been an immense source of happiness and reprieve from my work. Acting as a distraction (in the best and worst ways) and an incentive. They relieved my anxiety when I was feeling stressed and worried, and their love and admiration reassured me when I didn't feel I could go on. I am a better person with them in my life and for that I am grateful, I just hope that over the years, I can repay them for that.

I would also like to thank my PhD supervisors, Matt, Niels, Glen and Mark, who have provided advice and feedback over the years. They have stood by me when I have struggled, and were always understanding and empathetic.

My final thanks go to my partner, Rachel, and her beloved children, Kaiya and Willow. She has been an immense support through the final parts of my thesis and without her reassurance, I doubt I could have made it over the final hurdles.

Thank you all.

Blank Page

Introduction

In the current era of large datasets and complex data, the ability to utilise this information in pertinent ways is becoming more and more important. With this, the medical world is no different to any other scientific or technological field. The ability to focus data and funnel it towards a key goal is crucial. With that in mind, we must therefore develop not only the tools to create such avenues of advancement, but also the metrics with which to measure them.

Clinical Prediction modelling takes advantage of the world's medical datasets to guide clinicians to make accurate predictions for their patients. These models can be designed using a wide array of methodologies of all shapes and sizes and by essentially comparing patients to similar ones who have come before, we can make educated data-driven decisions about their futures.

Since human lives are transitive and ever changing, it makes sense to use a methodology that is also transitive. Multi-State Models analyse how long patients remain in discrete states. By combining this analysis with the field of prediction modelling, we are able to provide predictions of how long patients are likely to remain in a given state, and forecast their pathways through these systems.

The application of such models to specific clinical fields can assist in their uptake, and fields where clinical prediction models may be less utilised can benefit greatly from these novel approaches. The field of Nephrology concerns patients suffering from kidney disease (both acute and chronic) and therefore includes the diagnosis and prognosis of Chronic Kidney Disease patients and the prescription of their medical treatments such as Renal Replacement Therapy.

This thesis aims to unite these three topics to discuss the methodologies involved in creating a Multi-State Clinical Prediction Model for use in Chronic Kidney Disease patients on their journey towards and through Renal Replacement Therapy.

Chapter 1 consists of an overview of the current literature in these three key fields and discusses core concepts in each of them such as the design of a Clinical Prediction model, the statistical modelling involved in Multi-State Modelling and the factors of most importance when predicting Chronic Kidney Disease outcomes.

Chapter 2 dives into pragmatic approaches to measuring treatment effects. Particularly when Competing Risks and confounding effects are involved. Competing Risks are a subdivision of Multi-State Models and occur much more often in the real-world than in Randomised Controlled Trials. This is a simulation study investigating how Competing Risks can interfere with otherwise simple measurements.

Chapter 3 conducts another simulation study focusing on metrics involved in assessing the

calibration of a Clinical Prediction Model. In particular it compares the current best approaches to measuring calibration in time-to-event data where patients may be censored before having the event in question.

Chapter 4 builds on the simulation by applying similar adjustments to some other metrics for Clinical Prediction Models and extends these to a Multi-State Model framework by utilising similar work devoted to multinomial outcomes.

Chapter 5 culminates by developing and validating a Multi-State Clinical Prediction Model using two large datasets in disjoint populations. It predicts the outcomes for Chronic Kidney Disease patients, and their probability, over time, of commencing Renal Replacement Therapy and/or dying.

The thesis includes several appendices which augment the knowledge of the main text's chapters by providing supplementary material with additional discussions and in-depth investigations of some of the smaller topics.

Chapter 1

Literature Report

1.1 Introduction

This Chapter aims to discuss three themes central to the current research project:

- Prognostic Research
- Mult-State Models
- Chronic Kidney Disease

Section 1.2 discusses prognostic research from a general standpoint. The PROGRESS Series [1]–[2] classified prognostic research into four main categories, however the main focus of this Project has been the third of these, Prognostic Model Research[3] as it is the most relevant to the overarching project.

Section 1.3 will discuss Competing Risks (CRs) and Multi-State Models (MSMs), which are extensions of survival analysis. In ordinary survival analysis, patients can be in only one of two states, alive or dead (or their analogues), these extensions can include multiple death states (CR) and/or multiple living states (MSM) giving a more granular and in-depth assessment of a patients journey.

The final section, Section 1.4 will discuss methods of modality of Renal Replacement Therapy (RRT). RRT consists of three treatment methods designed to replace the functionality of the natural kidneys in patients.

1.2 Clinical Prediction Models

The idea of prognosis dates back to ancient Greece with the work of Hippocrates [4] and is derived from the Greek for "know before" meaning to forecast the future. Within the sphere of healthcare, it is defined as the risk of future health outcomes in patients, particularly patients with a certain disease or health condition. Prognosis allows clinicians to provide patients with a prediction of how their disease will progress and is usually given as a probability of having an event in a prespecified number of years. For example, QRISK3 [5] provides a probability

that a patient will have a heart attack or stroke in the next 10 years. Prognostic research encompasses any work which enhances the field of prognosis, whether through methodological advancements, field-specific prognostic modelling or educational material designed to improve general knowledge of prognosis. Prognostic models come under the wider umbrella of predictive models which also includes diagnostic models; because of this most of the keys points in the field or prognostic modelling can be applied to diagnostic models with little to no change.

Prognosis allows clinicians to evaluate the natural history of a patient (i.e. the course of a patient's future without any intervention) in order to establish the effect of screening for asymptomatic diseases (such as with mammograms [1]). Prognosis research can be used to develop new definitions of diseases, whether a redefinition of an existing disease (such as the extension to the definition of myocardial infarction to include non-fatal events [6]) or a previously unknown sub-type of a disease (such as Brugada syndrome as a type of cardiovascular disease [7])

In general, prognosis research can be broken down into four main categories, with three subcategories [8]:

- Type I: Fundamental prognosis research [1]
- Type II: Prognostic factor research [9]
- Type III: Prognostic model research [3]
  - Model development [10]
  - Model validation [11]
  - Model impact evaluation [12]
- Type IV: Stratified Medicine [2]

For a particular outcome, prognostic research will usually progress through these types, beginning with papers designed to evaluate overall prognosis within a whole population and then focusing in on more specificity and granularity towards individualised, causal predictions.

The model development and validation will usually occur in the same paper [13], [14]. studies into all three of the subcategories of prognostic model research *should* be completed before a model is used in clinical practice [15], although this does not always occur [3]. External validation is considered by some to be more important than the actual derivation of the model as it demonstrates generalisability of the model [16], whereas a model on it's own may be highly susceptible to overfitting [17].

1.2.1 Types I & IV

4. P 22 – a brief explanation of fundamental prognostic research and stratified medicine should be included

Fundamental Prognosis Research

Fundamental Prognosis Research is concerned with analysing the progress of a disease in the general population such as the overall survival rate of a disease [1]. This usually includes the application of standard care and can include a general level of stratification, such as investigations into the rate of survival in different countries (which can then be an indicator of how successful different healthcare regimes have been on disease progression). It is the building blocks needed to demonstrate evidence and a need for further research into specific areas.

A rather important aspect of fundamental prognosis research is that it allows other researchers to better understand the shape and scale of the progress of a disease. As well as providing insight into the average progress of a disease, it can also provide details of outliers and how the rates and differences are spread throughout a population, as well as indicators to what could be effecting the outcomes [18].

As the name suggests, this is a very broad area of research and encompasses many types of study that can be considered prognosis research, but that don't nicely fit into the other categories. It has also become a very prominent area in recent years with the analysis of the spread and survivability of COVID-19 being included in this category of research [19], and even the continuous updates of numbers of deaths in each country is a part of this field and the dashboards derived from them [20, 21].

Stratified Medicine

Once clinical prediction models have been created (through the techniques described below), they can be used to estimate which types of patients will have the highest risk of an outcome. This can inform clinical practice since those patients who are at high risk will benefit more from preventative treatment than those with low risk of an outcome [2]. This idea of using prognostic indicators to decide which patients should get a treatment is called Stratified Medicine.

In the UK, a common use of stratified medicine is through the use of the QRISK3 clinical prediction model [5], where clinicians are recommended to prescribe statins to prevent cardiovascular events for patients whose risk of an event in the next ten years is over 10% [22]. This also incorporates the decisions of which types of statin to prescribe depending on certain comorbidities due to the need to avoid certain interactions.

The actual research involved in stratified medicine is focusing on finding those areas where this kind of approach produces the most utility for patients and investigating which factors have the greatest effect on patient outcomes. It can also include analysing where treatments have different effects on patients depending on their characteristics, such as mutations causing medicines to work better [23] or worse [24] in certain conditions.

Causality

Causality is the study of which factors or covariates cause certain outcomes to occur [25]. This is the primary focus of studies such as randomised clinical trials, which attempt to establish whether changing a treatment. These can also be used when establishing the effectiveness of a clinical prediction model (see section 1.2.5).

The study of prognosis research is primarily concerned with prediction and not with aetiological research, that is investigating which covariates are considered to be causal. Clinical prediction research can lead to a deeper understanding of causality and aspects of this research can be used to confirm or refute the presence of causal relationships between covariates and outcomes, but is not intrinsic to this research.

Causal relationships often have to be taken into account when studying prediction as knowledge about what covariates *can* effect an outcome, can provide a guide on which covariates we should consider as part of our clinical prediction research. For example, we know that smoking status has causal relationships with many outcomes including a variety of cancers and vascular disease [26] and so it has become a key measurement taken in most research, including clinical prediction.

A side effect of how causal relationships can be studied within clinical modelling (in general) is that we can uncover covariates known as confounders. These are variables which have an effect on the outcome being measured, and which also has a relationship with the predictor being measured [27], see figure 1.1. When trying to establish a causal relationship between two variables (e.g. a predictor and an outcome), it is important to ensure that there isn't another (possibly hidden) relationship which may be causing the actual effect that is being observed. For example, older people tend to be more likely to be given prescriptions for certain medications. Since there is likely a relationship between age and propensity of medication use, then the age of the patient should always be taken into consideration when assessing the use of these medications and their effect on outcomes, especially in an observational setting.

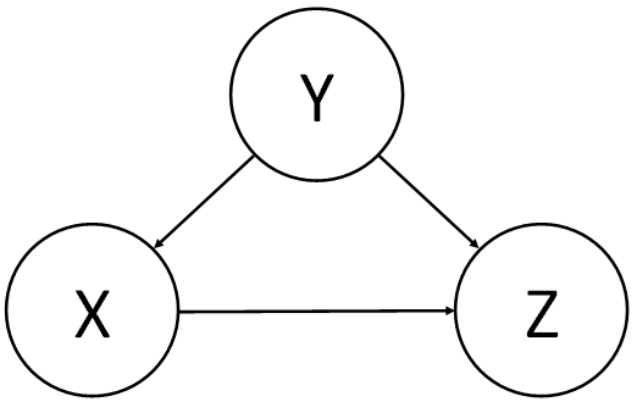

Figure 1.1: X has an effect on Z, Y has an effect on X and Z, therefore Y is a confounding factor on the relationship between X and Z

1.2.2 Prognostic Factor Research

The aim of prognostic factor research (Type II) is to discover which factors are associated with disease progression. This allows for the general attribution of relationships between predictors and clinical outcomes.

Prognostic factor research can give researchers and clinicians an idea of which patient factors

are important when assessing a disease. It is vital to the development of clinical predictive models as without an idea of what covariates are associated with an outcome, we cannot figure out which variables will useful for a prediction model.

For example, in the development of the NPI [28], it was known that cancer tumour grade is a prognostic factor in cancer patients as a higher grade correlates with a higher mortality. Note that the word correlate in the previous sentence. A recurring theme of statistics is the idea that correlation does not imply causation. Margarine consumption is correlated with divorce rates, Nicolas Cage films are correlated with pool drownings and chicken consumption is correlated with oil imports [29], but it is clear in all of these cases that there is no causal relationship between the variables. Predictive (and therefore prognostic) factor research is not aimed at discovering causal relationships, but merely uncovering correlations [12]. These factors may indeed be causal, but this is not a requirement [9]. In neonatal case, the Apgar Score is simple quick-to-use CPM for the health of a newborn [12]. In the Apgar score, skin colour (Blue all over, Blue at extremities or no cyanosis) is used as a prognostic factor, but it is not causal.

Counter to the idea that prognostic factors aren't always causal, they can also be confounding factors for the event they predict. Thus prognostic factors should be taken into account when planning clinical trials as if they are wildly misbalanced across the arms (or not accounted for in some other manner), they can cause biases in the results [9]. Sometimes these factors are so strong that adjusting the results of a clinical trial by the factor can affect, or even reverse the interpretation of the results [30]. If a prognostic factor is causal, then by directly affecting the factor, it can causally affect the outcome. By discovering new prognostic factors, and investigating their causality, we can potentially open the door to new directions of attack for treatments.

Some prognostic factors might only elicit a response for patients on treatment (such as metabolic effect removing a drug from the patient's system faster) and give no response to those without treatment [2]. For example, CYP2C9 and VKORC1 genotypes have an influence on patients being treated with warfarin, but have no effect on a patient's risk of stroke without treatment [31], [32]. This is called a differential response.

A positive result from aBRCA2 test can be an indicator of a risk of breast cancer amongst women [33]. If two women, one young and one old, both test positive for the gene, the young woman would likely be advised to have a mastectomy, whereas the older woman would not. This is because the younger woman is of higher risk of dying from breast cancer (even though breast cancer has not been diagnosed) than the older woman [31]. This indicates that, although the prognosis of the two patients would be similar, there are other factors, such as age and risk of other disease/cause of death, that need to be taken into account. The younger woman would have a greater decrease in relative risk of death than the older woman.

A prospective study by Fliser at al [34] found that Fibroblast Growth Factor 23 (FGF23) plasma concentrations was a prognostic factor for the progression of CKD. Haemoglobin A_1c, levels should be routinely measured in patients with diabetes as it is a prognostic factor for vascular events [35]. The Systemic Lupus International Collaborating Clinics (SLICC)/American College of Rheumatology (ACR) Damage Index (SDI) is used to measure the level of permanent damage in a patient with Systemic Lupus Erythematosus(SLE) [36]. Even though this is

a derived measure, rather than an explicit one, it is still a prognostic factor for cardiovascular related deaths [37]. Different ethnic groups can have vastly different prognosis in regards to cardiovascular disease[38], [39] and so ethnicity (and family history) are prognostic factors for most clinical events. These are all examples of prognostic factors from a wide variety of clinical sources.

It is unfortunate that, as discovered by Riley et al, only 35.5% of prognostic factor studies in paediatric oncology actually reported the size of the effect of the prognostic factor they reported on. This means that very little information can be drawn from these studies. It is also important that prognostic factor research papers consider and report on the implications of the factor they assess such as healthcare costs. These kinds of implications are rarely assessed, especially when compared to drugs or interventions [9].

1.2.3 Model Development

Prognostic factors can be combined into a prognostic model, which is a much more specific measurement of the effect of a factor on an outcome [3] and they are deigned to augment the job of a clinician; and not to completely replace them [12]. Diagnostic prediction model can be used to indicate whether a patient is likely to need further testing to establish the presence of a disease [13], [14]. Prognostic prediction models can be used to decide on further treatment for that patient, whether as a member of a certain risk group, or under a stratified medicine approach [13], [14]. Outcomes being assessed in a prediction model should be directly relevant to the patient (such as mortality) or have a direct causal relationship with something that is [12]. There is a trend of researchers focusing on areas of improvement that are of less significance to the patient than it is to a physician [40]. For example, older patient's might prefer to have an improved quality of life than an increase in life expectancy, and thus models should be developed to account for this.

Clinical predictive models can take a variety of forms, such as logistic regression, cox models or some kind of machine learning. Regardless of the specific model type being used, there are certain universal truths than should be held up during model development which will be discussed here. The size of the dataset being used is of vital importance as it can combat overfitting of the data, but so is choosing which prognostic factors to be included in the final model. This section will discuss various ideas that researchers need to account for when developing a model from any source and can be applied to any model type.

By considering a multivariable approach to prediction models (as opposed to a univariable one), researchers can consider different combinations of prognostic factors, usually referred to as potential predictors [9]. These can include factors where a direct relationship with the disease can be clearly seen, such as tumour size in the prediction of cancer mortality [28], or ones which could have a more general effect on overall health, such as socio-economic and ethnicity variables [41]. By ignoring any previous assumptions about a correlation between these potential predictors and the outcome of interest, we can cast a wider net in our analysis allowing us to catch relationships that might have otherwise been lost [42]. Prediction models should take into account as many prognostic factors as possible. Demographic data should also be included as these are often found to be confounding factors, variables such as ethnicity and

social deprivation risk exacerbating the existing inequality between groups [43].

This section will highlight some of the important factors involved in model development.

Population Size

When developing a predictive model, the size of the dataset being used in an important consideration. A typical "rule of thumb" is to have at least 10 events for every potential predictor [44], [45], know as the Events-per-Variable (EPV). Recently, this number has been superseded by a methods to evaluate a specific required sample size [46] based on Events-per-Predictor (EPP), where categorical variables are transformed into dummy variables prior to calculation (therefore number of predictors is higher than the number of variables). These predictions use the R^2 value of a model (or the expected R^2 value if the model is novel) and are designed to satisfy specific criteria related to model quality. If there aren't enough events to satisfy these criteria, then some potential predictors should be eliminated before any formal analysis takes place (for example using clinical knowledge) [47]. For the purposes of accurate validation, it is also recommended that this development dataset contain at least 100 events (regardless of number of potential predictors) [15], [48], [49]

A systematic review by Counsell et al [50] found that out of eighty-three prognostic models for acute stroke, less than 50% of them had more than 10 EPV. Having a low EPV can lead to overfitting of the model which is a concern associated with having a small data set. Overfitting leads to a worse prediction when the model is used on a new population which essentially makes the model useless [10]. However, just because a dataset is large does not imply that it will be a *good* dataset if the quality of the data is lacking [15]. Having a large amount of data can lead to predictors being considered statistically significant when in reality they only add a small amount of information to the model [15]. The size of the effect of a predictor should therefore be taken into account in the final model and, if beneficial, some predictors can be dropped at the final stage.

Large datasets can be used for both development and validation if an effective subset is chosen. This subset should not be random or data driven and should be decided before data analysis is begun [15]. Randomly splitting a dataset set into a training set (for development) and a testing set (for internal validation) can result in optimistic results in the validation process in the testing set. This is due to the random nature of the splitting causing the two populations to be too exchangeable, which is similar to the logic behind the splitting of patients in a Randomised Control Trial (RCT). Splitting the population by a specific characteristic (such as geographic location or time period) can result in a better internal validation [11], [51]. Derivation of the QRISK2 Score [52] (known later as QRISK2-2008) randomly assigned two thirds of practices to the derivation dataset and the remainder to the validation dataset. This model was further externally validated [53], and its most modern incarnation, QRISK3, performed the external validation in the same paper [5] The Nottingham Prognostic Index (NPI) was trained on the first 500 patients admitted to Nottingham City Hospital after the study began [28] and later validated on the next 320 patients to be admitted [54], this validation was not performed at the same time as the initial development and is thus an external validation.

If a sufficient amount of data is available and it has been taken from multiple sources

(practices, clinics or studies), then it should be clustered to account for heterogeneity across sources [55]. It is important that any sources of potential variability are identified (such as heterogeneity between centres) as this can have an impact on the results of any analysis [1], [15]. Heterogeneity is particularly high when using multiple countries as a source of data [56] or if a potential predictor is of a subjective nature, which leads to discrepancies between assessors [57]. Overlooking of this clustering can lead to incorrect inferences [55]. The generalisability of the sources of data should also be considered in the development of a model. For example, the inclusion and exclusion criteria of an RCT can greatly reduce generalisability if used as a data source [12].

Model Selection

A prediction model researcher needs to select clinically relevant potential predictors for use in the development of the model [10]. Once chosen, researchers need to be very specific about how these variables are treated. Any adjustments from the raw data should be reported in detail [13], [14]. Potential predictors with high levels of missingness should be excludes as this missingness can introduce bias [10]. One key fact that many experts agree on is that categorisation of continuous predictors should be avoided [30] as it retains much more predictive information. The cut-points of these categorisations lead to artificial jumps in the outcome risk [47]. It is also worth noting that cut-points are often either arbitrarily decided or data-driven with the latter leading to overfitting [47]. If categorisation is performed, clear rationale should be provided with an acknowledgement that this will reduce performance [16], [58]. When applying a model to a new population, extrapolation of a model should be avoided [59] and so to aid in this, the ranges of continuous variables, and the considered values of categorical variables should be reported [16]. this is especially true for age. QRISK2 was derived in a population ranging from 35 to 74 years of ages and so should not have been applied to patients out of this range [43]. This ranges was later extended with the updated version [60] and currently can be applied to patients aged 25-84 [5].

When building a prediction model, we begin with a certain pool of potential predictors and try to establish which to include in the final model [47]. With k candidate variables, we have 2^k possible choices which can get unwieldy even for low values of k, with only 10 predictors (a very reasonable number), there are over 1,000 combinations. This doesn't include interactions or non-linear components which increases this number even more. Therefore, model-building techniques are important for anybody attempting to build an accurate prediction model. It is currently undecided what the "best" way to select predictors in a multivariable model is or even if it exists [47]. One method that researchers use to decide on which predictors to include is to analyse each potential predictor individually for a correlation with the outcome in a univariable analysis and keeping those which are considered to have a statistically significant correlation. The general consensus amongst researchers is that predictors should not be excluded in this way [10]. Univariable analysis does not account for any dependencies between potential predictors and so any cross correlations that exists between them can cause a bias in the results since the 'best' values are not adjusted by the other values. This can result in variables being included that are not optimal, for example two variables are that highly correlated with each other may

be included, when a better model could only have one of them, which also results in a simpler model. Despite its clear weaknesses, many prognostic studies still use univariable analysis to build their models [61].

The NPI predictive model includes lymph-node stage, tumour size and pathological grade to identify patients with a poor prognosis with much better discrimination that would be possible if only one of these factors were used in isolation [28]. The development of the model began with nine potential predictors, of which three were considered to be statistically significant in a Cox model [62] and so were included in the final model which was simplified to $I = 0.2 \times \text{size (in cm)} + \text{stage} + \text{grade}$.

Backwards elimination (BE) involves starting with all potential predictors in the model and removing ones which do not reach a certain level of statistical significant (for example, 5%) one at a time until all remaining variables are significant. Forward selection begins with no variables and adds one a a time based on similar criteria. Under either of these methods, a lower significance level will exclude more variables [10]. Backward elimination of variables is preferable over forward selection as users are less likely to end up in local minima and is more likely to produce robust models because it considers the effect of all variables initially and thus helps to avoid collinearity [63]. A variant of these techniques is to use the Akaike Information Criteria (AIC) rather than statistical significance. This method avoids the comparison to p-values and so is often preferable to build robust models [64]. For this method, to establish which predictors should be removed at each step, the model is re-built with each of the predictors individually removed, and the AIC is calculated. The model with the lowest AIC is chosen to be the new model and the process is repeated. This process is repeated until the removal of a predictor would increase the AIC (i.e. make the model's fit worse). This same technique can be applied to a forward selection style model or, if the computing power is available, a backward-forward elimination technique were predictors are added or removed at each stage. The advantage of this method is that it avoids local minima better by trying more combinations.

It is also important to assess non-linearity relationships between variables and outcomes to ensure the relationship is accurately modelled. This can be done using standard transformations (e.g. logarithms, squaring) or using fractional polynomials [65]. Interactions between terms also need to be checked for the same reasons, and when interactions are strong, it may be useful to completely stratify by a factor, rather than including as a covariate in the model. Strong interactions can be an indicator for a differential response amongst populations and so should be investigated directly [2]. If a predictor is expensive or invasive, it may be better to include a less significant predictor which is easier to come by [12]. A limiting factor for some prognostic models is that the prognostic factors they measure are not readily available or are not used in routine care [3]. The measurement (or lack thereof) can also be an indicator of patient health and so researchers need to be aware of these causal links when analysing measurements [66], this is similar to the idea of data being not missing at random as discussed below.

When dealing with time-to-event data, it is often not the case that the effect of a variable is constant over time. The assumption of the naïve Cox Model [62] is that the effects don't change over time (whether linear or not), however much like alternative methods, such as the Royston-Parmar model, the Cox model has been extended to take these time dependencies into

account [67].

QRISK2 checked for non-linearity amongst continuous potential predictors using fractional polynomials as well as certain interaction terms (in particular interactions with age) [43], [65]. Where an interaction in factors is identified in a study, this can be a useful indicator of a differential response and should be investigated further [2]. If a predictor is expensive or invasive to measure, it might be better to include a less significant predictor which is easier to come by [12]. A limiting factor for some prognostic models is that the prognostic factors they measure are not readily available or are not used in routine care [3]. Sometimes, the measures may be partly known and partly unmeasured as is the case with missing data.

Missing Data

During development of any model, researchers should assess the level of missingness within their data and investigate what kind of missingness is present [68]. Missing data is usually classified in three ways:

- **Missing Completely At Random (MCAR)** - Whether a piece of data is missing is purely by chance and is not linked to any other factor (whether measured or unmeasured). If data is MCAR, then the distribution of the missing values will be approximately the same as those that are present.

- **Missing At Random (MAR)** - There is an underlying link between whether a piece of data is missing, and other factors, however, we have measurements for all of the other factors. For example, we measure both X & Y, and we know that Y is more likely to be missing when X is high. Missing values will thus follow a different distribution pattern than those that are measured, however we can often estimate this based on the data that is available.

- **Missing Not At Random (MNAR)** - There is an underlying link, as with MAR, however, the variables that cause the missingness are not measured within our data and thus we cannot easily produce an effective model for the missing data.

If it is believed that the data is MCAR, then the use of complete case analysis (only patients for whom every item of data is available) can be justified, alternatively if only when a small percentage of patients ($< 5\%$) do not have all available data, it can also be acceptable [69]. However if this is not the case, it could vastly reduce the power of the results due to the lower number of patients in the population and if the assumption of MCAR is incorrect, it can induce a bias in results [10]. Usually, however, data will be missing due to some underlying relationship. This relationship may be due to measured or unmeasured confounders. If the confounder is measured, then the data is MAR, if it is unmeasured, then the data is MNAR, or it may be a combination of the two, in which case it would still be considered MNAR. Using causal graphs can assist researchers in determining which type is present by investigating the causal relationships at play [70]. A causal graph, presented as a Directed Acyclic Graph (DAG) allows researchers to determine which variables might have an effect on the others [71], [72]. These are usually designed with a vast amount of clinical knowledge to influence the choices of

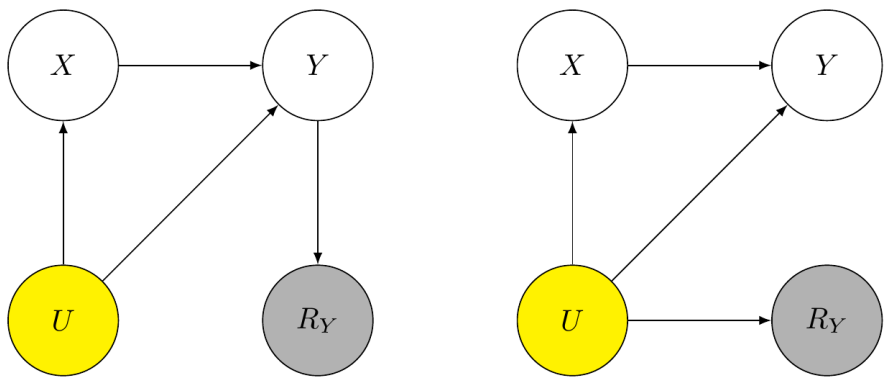

Figure 1.2: Examples of Missing Data Causal Mechanisms. The left shows data that is MAR, missingness is influenced by Y and the right shows data that is MNAR, missingness is influenced by U, an unmeasured variable

connections. The missingness of a variable can also be included as a node on a causal graph [73]. Each variable is represented by a node, and the direction of causation between two variables is an arrow. In Figure 1.2, X has a causal effect on Y and U has a causal effect on X, where U is an unmeasured variable.

Just as with regular causal systems, researchers need to determine what mechanisms could cause the data to be missing before filling it in with estimates. This can be done through a method called Multiple Imputation (MI), where the missing data is estimated multiple times with an element of random-ness each time[74]. Analysis is performed on each of the datasets in turn, and the results are aggregated [68]. There are many ways to perform MI, the most commonly used is called MICE, which stands for Multiple Imputation Chained Equations, where the missing data are initially estimated, and then re-estimated iteratively [31]. If the causality of missingness is linked to a survival related variable, it is recommended to include the Nelson-Aalen estimate of survival (see section 1.3.1) [75].

During development of the NPI, only 387 of the 500 patients were included in the study due to missing data in the other 113 [28], however QRISK2 did use multiple imputation methods to deal with the missing data [43]. In a review of CKD predictive models by Collins et al[16], out of eleven models, four conducted complete case analysis only and only two conducted multiple imputations on the missing data. The remaining five studies did not mention missing data at all. Small data sets, inappropriate handling of missing data and lack of validation are common issues in prediction model development [13], [14].

Presentation

It is often difficult to convey information developed through scientific research, particularly to non-scientists [76]. The field of clinical prediction modelling is no different. To quell these problems, Bonnett et al [77] created an informative guide to presenting prognostic research to clinicians.

Once developed, prognostic models can be used to create risk groups for a population. Risk groups should be defined by clinical knowledge rather than statistical criteria [11]. Grouping

patients into risk groups is not as accurate as using the specific model to provide an estimated risk [3], but can often be more easily understood by users or the model. The original development paper for NPI,patients were classified into three risk groups, Low (I < 3.4), Medium (3.4 < I < 5.4) and High (I > 5.4). The follow-up paper extended these groups to be: Very Good (I < 3), Good (3 < I < 4), Moderate (4 < I < 5), Poor(5 < I <6) and Very Poor (I > 6) with annual percentage mortality rates of 1.5, 3.5, 6, 20 and 32 respectively.

Models can be simplified to provide a point scoring system [31], which often requires the removal of less significant covariates, and the rounding of model parameter estimates. The advantage of this is that models can be easily implemented, which can be necessary when time is of the essence, such as in emergency situations [31], [78] . These simplified models can often be transformed into a graphical chart to be even more intuitive.

As well as predictive models, more subjective methods of assessing a patients life expectancy have been proposed, such as Moss et al[79] who suggested the physician asks themselves *"Would I be surprised if this patient dies in the next year?"* to identify high mortality amongst dialysis patients. However these types of prediction should be used with caution as they are dependent on the physician in question and are thus subjective analyses and if used as a potential predictor in a model, would be highly susceptible to clustering effects.

A very small number of developed models are currently routinely used in clinical practice[16], however their use is becoming more common, with more and more healthcare providers recommending their use?[13], [14]. This current lack of clinical implementation may be due to a lack of standardisation and inadequate reporting. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [13], [14] was developed in an attempt to combat this inconsistency. It is described as *"a guideline specifically designed for the reporting of studies developing or validating a multivariable prediction model, whether for diagnostic or prognostic purposes"* [13], [14].

As acknowledged in the QRISK2 development paper [43], as new technologies arise (such as eHealth Records (EHRs)and genomic measurements), it is important to update existing models with added information. Models should be constantly updated and re-validated as populations and ambient health effects change over time [3]. Clinicians and researchers should be wary of models becoming outdated [80]. Healthcare systems and lifestyles change over time, and so models developed and externally validated in an outdated population will drift [81] and so should be updated regularly, as with QRISK [43] and it's follow-up implementations [5], [60] or automatically with a dynamic model [82].

Prime Examples

Since QRISK2 has been developed in association with Egton Medical Information System (EMIS), there is the ability to automatically apply a QRISK2 measurement to all patients in EMIS and thus produce a list ordered by risk score to establish which patients are most likely to suffer a cardiovascular event in the next 10-years and to which patients resources can be prioritised [43]. Because of this, during development, it was estimated that QRISK2 would be generalisable to around 80% of practices in the UK although they acknowledged that the model should still undergo external validation [43]. Derivation of QRISK2 was not done on

patients with a history of cardiovascular disease and so it cannot be applied to them. It was acknowledged in the original QRISK2 development paper that it would require updating, which now happens annually with more up-to-date data [83], and eventually called for an overhaul with QRISK3 @ [5].

The systematic review by Counsell et al [50] investigating acute stroke found that only four of the eighty-three clinical prediction models met their 8 key quality criteria. Models were assessed on the following:

- Externally Validated
- Adequate inception cohort (i.e. large enough EPV)
- Less than 10% loss to follow-up
- Prospective data collection
- Valid and reliable outcome
- Age as a candidate predictor
- Severity of condition as a candidate predictor
- Use of stepwise regression

None of the four models which satisfied the last seven criteria had been externally validated, and thus none of the models satisfied all eight criteria. Over 150 different predictors were considered across the review with most of them only occurring in only one or two of the modes. Of the 18 variables that were assessed in at least five models, only three were significant in more than 80% of the relevant models. This demonstrates a lack of cohesion between research groups with a disregard for previously developed models and a preference for researchers to use available data sources to develop new models rather than update existing ones [3], [11].

In their systematic review of CPMs in CKD, Collins et al [16] assessed ten study papers. In this review, they identified 97 potential predictors that were considered across the ten studies, with 58 of them only being considered in a single study with a median of four potential predictors per study and only six of the studies justified their choice of potential predictors. Only three of these studies reported the age range of the patients involved and only two studies assessed linearity in the predictors. A review of 47 papers by Mallett at al [84] found that the reporting of prediction model development in cancer was very poor in all measures. Another review of 71 papers by Bouwmeester et al [85] found that even the reporting of prediction model development in high-impact medical journals was incredibly poor.

1.2.4 Model Validation

Creating a clinically useful model is not as simple as just using some available data to develop a model, despite what a lot of researchers seem to believe [86]. To quote Steyerberg et al [3]: *"To be useful for clinicians a prognostic model needs to provide validated and accurate predictions and to improve patient outcomes and cost-effectiveness of care"*. This means that, although

a model might appear to be useful, its effectiveness is only relevant to the population it was developed in.

There are two main types of validation, internal and external. Internal validation uses the same dataset as the model was developed in, whereas external validation uses a novel dataset and is often done as a follow-up project and usually incites a second paper (possibly by a new team of researchers). There are various methods for internal validation which will be discussed in this section as well as the advantages that external validations possess over internal ones. The validity of a model can be measured in various ways which can usually be classified as either a discriminatory measure, a calibration measure or an overall accuracy measure and all measurements can be applied to both internal and external datasets.

The external validation of a prognostic model is considered by some to be more important than its development as it demonstrates the generalisability of the model, without which, a prediction model is essentially useless [3]. Because of this, the TRIPOD guidelines strongly recommends researchers to perform an external validation on their models, whether as part of the initial development paper or a subsequent one [13], [14]. This also means that unvalidated models should not be used in clinical practice [11], [87]. Despite this, they are often accepted as they are without being rigorously tested [87]. This means that clinicians should be wary when using predictions from models that are yet to be externally validated [12], however, even models which perform moderately under external validation are likely to be better than a subjective assessment [51]. Unfortunately, external validation studies are scarce, especially when compared to development studies [11], [88]. Hopefully, the ability to access big datasets such as EHR or Individual Participant Data (IPD) will allow external validation studies to blossom in the coming years [15]. For example, the QRESEARCH database, which was used to create the QRISK and QKIDNEY scores database contains 24 million patients from 1300 general practices[89]. The database contains longitudinal, demographic and mortality data on the individual patient level as well as many other factors that can be used for clinical prediction [43].

Discrimination

Discrimination is the ability of the model to separate out patients who are more likely to have an event from those that are less likely [11], [90], [91].

The D-statistic is a common measure of discrimination for time-to-event data. To find the D-statistic, we split the validation dataset into two at the median value for the prognostic index. The D-statistic is then the log hazard ratio between these two groups [15]. A higher D-statistic indicate greater discrimination [15].

The c-statistic is the probability that if we choose two patients at random, one who had the outcome and one who did not, the patient who had the event will have been predicted a higher risk than the patient who did not have the event [15]. When extended to time-to-event data, this translates to patients with higher scores having the event *earlier* [31]. For binary outcomes, it is known as the area under the Receiver Operating Characteristic (ROC) [15], [87]. Higher values of the c-statistics implies better discrimination [43]. The c-statistic is between 0 and 1 with 1 indicating perfect discrimination and 0.5 indicating that the model is no better

than chance, and therefore non-informative [15]. If in the binary outcome case, a c-statistic is between 0 and 0.5, the results from the model can be reversed to provide a model with a $1-c$ c-statistic (e.g. if a model has a c-statistics of 0.25, by reversing the outcome, you create a model with a theoretical c-statistic of 0.75, however this new model would have to be externally validated). The c-statistic has been criticised for its inability to detect meaningful differences [92] and is commonly between 0.6 and 0.85 for prognostic models [12].

The Framingham Score, which is used to predict risk of cardiovascular event, was found to have a c-statistic of around 0.70 [93]. Fraccaro et al [94] conducted a study using a large EHR dataset(n = 178, 399) to validate seven models that predict the onset of CKD and found that all seven had a c-statistics of around 0.9 indicating high levels of discrimination.

Calibration

Calibration is the ability of a model to accurately predict the number of events in a group of patients (e.g. among patients who are predicted a 10% chance of having an event, 10% of them should have the event)[11], [90], [91]. The simplest way to assess calibration in a binary case is through an E/O calculation. The E/O calculation is simply the ratio between the average Expected probability within a population (or sub-population), and the proportion of patients who had the event, i.e Expected divided by Observed. In a perfect model, the E/O would be exactly 1 [15], if the E/O is above 1, then the model is, on average, over-predicting since this implies that average Expected is greater than the actually Observed events; similarly, if E/O is less than 1, then our model is under-predicting.

A more nuanced (but naïve) method to assess calibration is to group patients at tenths of predicted risks by their predicted (or expected) risk and plotting the average expected risk against the observed outcome within that group (i.e. the percentage of patients in that grouping who had the event)[11], [15]. A line-of-best-fit can be added to this plot which demonstrates the calibration of the model.

If the model is perfectly calibrated, then the points should all lie on the direct diagonal (Observed = Expected), and thus the line of best fit should have an intercept of 0 and a slope of 1 [15]. These measures are known simply as the Calibration Intercept and the Calibration Slope. We would usually derive the intercept with the slope fixed at 1, and then fix the intercept to this value and estimate the slope. In this case, the intercept is usually referred to as the Calibration-in-the-Large (CITL). If the CITL is below 0, then the model is systematically over-predicting patient outcomes, if it is above 0, then the model is systematically under-predicting patient outcomes. The Calibration Slope is below 1, then then some of the predictions are too extreme, whereas if it is more than 1, the predictions are too modest.

The most widely used method of calibration assessment, takes this idea and rather than grouping results, we perform a logistic regression on the Observed vs Expected results. It is important to note that this is applied to the raw data, and so no grouping occurs. This logistic regression gives us an intercept and slope measurement, which are similarly interpreted, but are no longer dependent on the size of subgroups we choose to plot [8]. We can also still plot this as a line-of-best-fit through the grouped calibration plot, however it will now be curved as a logistic curve, rather than a straight line.

Further methods to assess calibration are through the use of graphical calibration curves and the integrated calibration index (ICI) as per the recent work of Austin et al [95], wherein they apply a hazard regression model (such as an Royston Parmer model as discussed in 1.3.1) to the predicted probabilities and assess the outcome of the calibration model to establish the quality of a model graphically.

It should be borne in mind that summarising validation statistics is not always adequate and that assessing the validation of a model on different subgroups is more beneficial [15]. The calibration slope of all three versions of QRISK2 were assessed in male/female subgroups and ranged between 0.92-0.95 for men and women which indicates a very good calibration in *both* groups [31]. However, the Framingham Score has an E/O of 1.03 over an entire population, but in women aged between 40 to 64, it over-predicts and in women aged 70-74, it under-predicts [31].

Internal Validation

Internal validation used to assess the levels of overfitting, optimism and miscalibration of a model[13], [14]. Internal Validation is the application of the above methods to a population derived from the development population, in contrast to External Validation, which uses a different population. The ability to translate a model into a new population is called its generalisability. Because it uses the development data set, internal validation provides no information on the generalisability of a model, but merely indicates how well the model is calibrated for and discriminates in the development population. Although poor internal performance can imply poor generalisability, good internal performance does not imply good generalisability. Models can often be overfit, meaning they perform better on the data for which they have been developed than on external data [87].

Three common methods for performing internal validation are sample splitting, bootstrapping and cross-validation.

Cross-validation involves randomly splitting the dataset into m subsets (e.g. in 10-fold cross-validation, you have 10 subsets) and then developing a model on $m-1$ sets and then validating it on the other set. This is repeated m times, once for each data set, giving m new models which can be compared to the model produced from the entire dataset.

When bootstrapping, if we have a population of size n, we randomly select (with replacement) n patients from the population and validate on this new population. This is repeated a prespecified number of times (e.g. 1000 times) to give validation measures. Model optimism can also be estimated by taking the difference between the performance of the model in the bootstrapped datasets and the performance in the original development set [87].

The final method used for validation is sample-splitting where model developers randomly split their population into two subsets, a training set and a testing set, usually in a 2:1 ratio. The training set is used to develop the actual model which can then be validated against the testing set. Both cross-validation and bootstrapping provide sets of values for each of the validation measures we are using which can be aggregated to provide robust estimates (e.g. means and confidence intervals). These estimates would only be relevant to an external population which is extremely similar to the development dataset.

Bootstrapping techniques are considered to be an effective and efficient method for internally validating a dataset against the current population(as opposed to cross-validation or split-sample methods) [87]. Bootstrapping provides a nearly unbiased estimate of the effect of the predictive accuracy of a model [90], it can provide a shrinkage factor to partially re-calibrate a model to improve performance [87]. However, bootstrapping does not change the underlying population and so is not quite thorough enough and so external validation is still more important to the usability of the model than internal validation [96], unless the target population is similar enough to the development population.

External Validation

It is preferable that a model undergo External Validation before its use in the real world [3]. It is expected that models will perform worse in external validation than internal validation. However, if the performance is significantly worse, this may render the model essentially useless [87]. Different settings, both geographic and temporal can cause poor performance in external validation, for example, models developed in primary care are different from those developed in secondary care due to the higher rate of more severe conditions in secondary care[97].

When applying a model to a new dataset, it is important to compare the populations of the derivation dataset and the new dataset to ensure compatibility of the model[87]. This is done through a comparison of the case mix of a population. The case mix is the defined as the distribution of all of the potential predictor variables and the outcome variables (including those which do not make it into the final model) [12]. If these are similar between two populations, then we can assume the populations as a whole are relatively similar. Models will be more applicable to a new population if the case mixes of the two populations are similar [12]. Applying a prediction model to a similar population can improve the chances of the model fitting the dataset well, but this will reduce the applicability of the model to other populations in the future and might cause unwarranted confidence in the predictive abilities of the model.

High heterogeneity in the development population is an advantage as low heterogeneity leads to low discrimination (if patients are similar, their prognosis would also be similar) which can therefore reduce the generalisability of the model [15]. It would be difficult to ascertain how often this happens as external validation studies which results in poor performance are rarely published [98], [99]. If performance is not consistent across populations then users should be made aware of this and might even be advised to use a different model in certain cases [15]. Validating models developed in one setting in another setting is useful, but is expected to produce less than ideal validation statistics (discrimination and calibration) [15].

If, during external validation, a model does not reach a high enough standard of calibration, but has good discriminative ability, it is more useful to recalibrate the model (possibly by introducing a multiplicative factor) than to develop an entirely new model. By doing this, we are using the data from both the original development, and the new recalibration dataset, whereas if we were to abandon the old model, we would only be using a single dataset of data [100].

The recalibrated model may need to be revalidated in another population [12]. In a regression based model, recalibration may involve simply adjusting the coefficients of the model to

achieve a calibration slope closer to 1 [12]. If simple recalibration is not possible, updating the model in other ways is preferable over creating a new one and there are many techniques which can combine the original model with the new dataset [101]. However a model is updated, the new model should always be re-assessed for external validity [12].

Model development often takes precedence over model validation as a primary goal of research and most models don't make it out of the development stage of research [3]. It would be more productive for the field of prognostic models (and the specific fields for which the models are relevant) if there was a concerted effort to validate models more often rather than constantly developing new ones addressing the same questions [3]. This drive for "novelty" can lead to an influx of models which do not perform significantly better than each other and, ironically, cause a mass stagnation [102]. A large exception would be in developing methodologically driven models, which answer new questions.

Unfortunately, external validation studies will often be abandoned if poor performance is found, an idea that perforates the entirety of science through the hesitation to publish insignificant findings and a lack of replication studies [98]. Once an external validation study has been abandoned, researchers will then likely develop an entirely new model from their datasets and not publish the poor validation results. This means that rather than having a developed and validated model (which was the original intention of the project), we now have two unvalidated models for the same (or similar) outcome. This might cause another team to try and validate the original model (or the new model), leading to the same issue ad infinitum.

Prime Examples

During the original development, Hippisley-Cox et al used sample splitting techniques for the internal validation [43] and compared the performance of QRISK2 with the modified Framingham Score [103]. For this internal validation, patients were deemed as "High Risk" if they were predicted a 10-year risk of cardiovascular event of > 20 for each score separately. They then compared patients who were considered High Risk under one measure and not under the other to determine how a change in model usage (e.g. from Framingham to QRISK2) would impact patients. Overall, of the patients in the QRISK2 High Risk set, 23.3 of them had an event over the 10 year observation period and in the Framingham High Risk set, 16.6 had an event, indicating an underestimation by the modified Framingham compared to an overestimation by QRISK2. Amongst the two groups of reclassified patients, those in the QRISK2 High Risk set had an annual incidence rate of 30.6 and 35.2 for men and women,respectively. This is in contrast to the Framingham High risk set which had 25.7 and 26.4. This indicates that, at the 20 threshold, QRISK2 identified a more at risk population than Framingham did.

In 2012, Collins et al externally validated three versions of QRISK2 (from 2008, 2010 and 2011)[43], [60]. They used the The Health Improvement Network (THIN) cohort and each patient was given a predicted risk score based on the three QRISK2 equations as well as a fourth score based on a modification of the Framingham score [[103]; #97; #98].The regular Framingham Score calculates a risk for coronary heart attack and a risk for stroke separately, the modified score sums these values together and applies specific multiplicative effects depending on patient characteristics. Since the two risk scores are not necessarily independent this can

result in a higher than 100% risk for some patients. When this analysis was performed, QRISK2-2011 was the current version of QRISK, it has now been updated to QRISK3 [5]. When the QRISK2 model was developed, the National Institute for Health and Clinical Excellence (NICE) recommendations were to use the modified Framingham Score to assess a patient's coronary risk, however by the time the external validation was done, NICE recommended that physicians choose between Framingham and QRISK2 when performing this assessment based on their own experience with the two models [31].

Multiple imputation was used by Collins et al[31] to deal with missing data for the THIN cohort[104]. Amongst patients in the THIN cohort, the modified Framingham score over-predicted for most patients giving a very shallow calibration slope [31]. Most patients QRISK2-2011 scores were very similar to their QRISK2-2008 and QRISK2-2010 scores with almost all of them having their updated scores being within 3% of the older versions [31]. However in the internal validation, the calibration and discrimination were only *summarised* across all 176 practices which ignored potentially large heterogeneity in the population from practice to practice [15], [43].

A similar thing occurred in the external validation which ignored between-practice heterogeneity ($I^2 = 80.9$)[15], [31]. The external validation paper reported a 95% confidence interval for the c-statistic of 0.826 to 0.833 using data across 364 practices. However due to this between-practice heterogeneity, if the process were repeated with a new practice, we would predict it having a 95% confidence interval of 0.76 to 0.88. The conclusion of the external validation paper was for NICE to recommend that healthcare professionals abandon the Framingham score in favour of the QRISK2-2011 model, or at the very least to use a recalibrated version of the Framingham score[31]. Due to the improved predictive ability of the QRISK and QRISK2 scores and the results by Collins et al, the recommendation by NICE to use the modified Framingham Score was withdrawn in 2014 and clinicians are now advised to use the most recent version of the QRISK score available [31], [83] [5], [105].

For the validation study of NPI [54], it was assessed prospectively in a group of 320 patients at the same hospital as used for the development [28]. All 707 patients (387 original and 320 new) in the NPI validation study [54] were assessed by the same pathologist and under the care of the same surgeon as the development study [28]. As well as the three factors in the NPI model [28], the validation study also collected data on menopause status and Oestrogen Receptor (ER) as these were close to significant in the original study ($Z > 1.5, p < 0.134$). It is demonstrated graphically in the NPI validation study [54] that the risk groups produced in the original study are viable in the new population. The Cox analysis [62] was also re-run on the original 387 patients as more follow-up time was now available and the resulting coefficients were similar to the original study indicating the original model performs well in the extended dataset. This allowed Todd et al [54] to update the NPI risk groups to include five groups rather than the original three which allows for better stratification as mentioned earlier.

Fraccaro et al [94] performed an external validation of multiple models designed to predict the onset of CKD. It was described as "The first comprehensive head-to-head comparison study of multiple CKD prediction models on a large independent population" [94]. Five of the models assessed needed to be recalibrated to suit the population of Salford, UK. QKidney [106], which

was developed in the UK performed the best.

The model developed by Bleeker et al [87] to predict cause in febrile children failed during external validation. The ineffectiveness of Bleeker's model was confirmed by refitting a multivariable model to the new dataset which provided significantly different estimates for the coefficients. This was, however, useful as some of the predictors were still considered to have an effect on the outcome prediction, just at different magnitudes from the original, which implies that they can be useful as potential predictors in future models. These results stand as a testament to the necessity of external validation [87].

1.2.5 Impact Evaluation

As predictive models are considered to be health technologies, once they been developed and validated with good performance metrics, their usefulness in the real world will need to be assessed [3].This is done through an impact study. Although models with near perfect discrimination and calibration may not need an impact study to evaluate their usefulness [12]. From a health economics standpoint, it is useless to implement a model which costs more for the healthcare service than the current standard unless the improvements to population health are substantial. Impact studies should *always* be performed before a model gets used in clinical practice as they assess a models robustness and generalisability more thoroughly than an external validation [15], [87].

The impact of a prognostic model can be assessed in a manner similar to that of an RCT where some patients are provided treatment based on decisions made from a prognostic model and the control arm receives the best alternative treatment [3], [12]. Impact studies can be performed as a before and after comparison which may be simpler to implement, but can be sensitive to ambient changes in the clinical environment [107], [108].

Randomisation in an impact trial can happen at the patient, doctor or centre level. Higher levels (centres) are preferable over lower ones (patients) as this lowers the likelihood of cross contamination of the intervention (such as a single doctor treating patients on both arms of a study, or two doctors in the same centre discussing results)[109]. The intervention of an impact study can be an assistive or a decisive approach [12]. Assistive approaches provide the clinician with the patients risk score as indicated by the model and allows the clinician to use his or her own judgment on how to proceed with treatment. Decisive approaches explicitly tells the clinician what decision should be made based on the model (e.g. whether or not to prescribe statins). The STarTBack trial [110] was an impact study which compared the use of stratified care with traditional best care on patients with lower back pain. The results showed that the stratified care reduced the risk of disability and lowered the cost of care compared to the control arm. The NPI is widely cited and used, however its impact has hardly been assessed [3], [28].

1.3 Competing Risks & Multi-State Models

Many diseases are measured in stages of progression or as types or variants. Often, patients can switch from one of these stages to another whilst they are being studied. If being in a different stage of the disease is believed to affect the way that the patient's condition behaves

then it is important to account for this when modelling a disease. The simplest way is to have the disease stage/type as a covariate and ensure that it is updated accurately. However, if it is believed that a disease behaves wildly differently when at different stages, then this might not be feasible, especially if the stage can interact with other covariates. The solution to this is to use MSMs to map patients progression through the different stages of the disease [111], where each stage is modelled as a state in the MSM.

1.3.1 Traditional Survival Analysis

From survival analysis, a hazard function is a measure of the intensity of moving from one state to another (whether that is from alive to death, functioning to non-functioning or something more complicated as in an MSM). If we have T be the random variable defining the time of the event (or transition), then a hazard function is usually defined as [112], [113]

$$\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t) = \lim_{\Delta t \to 0} \frac{\mathrm{Prob}\,(T \leq t + \Delta T \,|\, T \geq t)}{\Delta t}$$

where S is the survival function, or cumulative probability of having remained in the current state from time $t = 0$. An alternative way of writing this is

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right)$$

We can simplify this equation to $S(t) = \exp\left(\Lambda(t)\right)$ if we defined the cumulative hazard function, $\Lambda(t)$, to be

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u$$

Two other useful definitions from survival analysis are the probability density function, f, and the cumulative distribution function, F which are much more familiar to statisticians and are related to the previously defined functions by:

$$f(t) = \frac{\lambda(t)}{S(t)} \qquad\qquad F(t) = 1 - S(t)$$

A function that is useful in estimating the survival of a population is the Kaplan-Meier estimate, it is a non-parametrics, empirical estimate of survival and is defined as:

$$\hat{S}(t) = \prod_{j:\, t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

In this definition, t_j is the jth event time, n_j is the number of patients still at risk at time t_j (i.e those event-free at time this time) and d_j is the number of patients who had an event at time t_j. Kaplan-Meier estimates assume independence between the event we are modelling and censoring [113]. See figure 1.3 for a typical K-M plot for two populations:

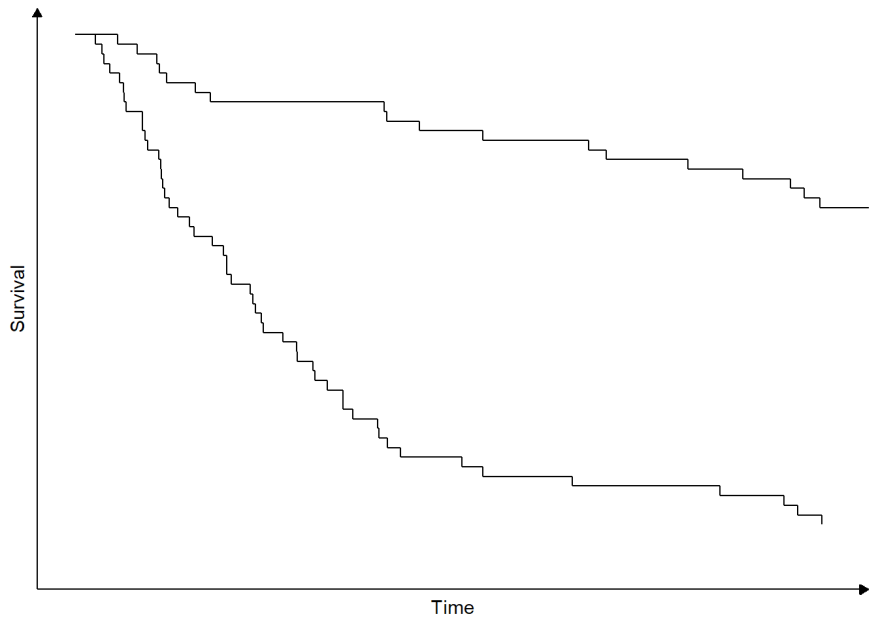

Figure 1.3: Example plot of Kaplan-Meier estimator for two populations

There are many different kinds of statistical models that can be used to produce clinical prediction models based on the type of data and the shape of the desired output. Many models rely on regression techniques to produce their estimates and these can usually be rearranged into a linear relations of the form:

$$\mathbf{Y} = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_m Z_m = \boldsymbol{\beta}^T \mathbf{Z} + \epsilon$$

where the $\boldsymbol{\beta}$s are the coefficients found from the data and the $\mathbf{Z}$s are the covariates of predictors. The first coefficient, β_0 is known as the intercept term and gives an idea of the average amongst the population (if the other covariates have been standardised). The final term here, ϵ is the error term or the residual, when measured across every patient, this error term should follow a Normal distribution and have its standard deviation be as small as possible, $\epsilon \sim N(0, \sigma)$. The ϵ term will be omitted from further equations, unless required.

The predictors above do not have to be directly from the raw data and can be derived in some way from the data (including via other regression models), and the predicted value here can be transformed by a link functions, usually called g to the actual expected outcome. By choosing the correct transformations and link functions and repeatedly applying these regression, we can form a simple machine learning model [114]. For example, a logistic model uses the logit function as a link function as seen below:

$$\text{logit}(p|\mathbf{Z}) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_m Z_m$$

The logit function can then be undone to provide a probability that an outcome occurs, such as the probability of a specific prognosis.

Within the realm of survival analysis, we have an extra dimension to include in our calculations, time, and this includes the fact that some patients are not observed after certain dates (i.e censoring). To combat this additional dimension, the most common form of regression in survival analysis, the Cox model [62] avoids estimating the intercept altogether and produces proportional hazard estimates for each covariate. Intercept values (i.e. the baseline hazard function) can be estimated for the Cox model using various techniques such as the Breslow estimator [115]. The predicted values from the model are positioned within the hazard function, described above and give an idea of how a covariate increases or decreases the hazard of an event:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp\left(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_m Z_m\right)$$

Notice that the intercept has, essentially been swallowed up by the $\lambda_0(t)$ term, which is known as the baseline hazard. This baseline hazard function is not estimated within the Cox modelling regression method and is assumed to be the same for all patients (subject to stratification). This means that estimated values are relative to one another and therefore absolute estimates are not possible with the Cox Model. These relative estimates, written in their exponentiated form, $\exp(\beta_j Z_j)$ are known as hazard ratios and are used extensively in clinical trials to compare two groups [116].

To provide any absolute estimates of hazard functions, we need to extend the Cox model and a reliable method to do this is with the Royston-Parmar Regression technique [67]. This method estimates the hazard ratios in much the same way as the Cox model does, but in the same process, also uses restricted cubic splines on $x = \log(t)$ [117] to estimate the log of the cumulative baseline hazard function, or the log-log of the baseline Survival function;

$$\log(-\log(S_0(t|\mathbf{Z}))) = \log(\Lambda_0(t|\mathbf{Z})) = \gamma_0 + \gamma_1 x + \gamma_2 \nu_1(x) + ... + \gamma_{m+1}\nu_m(x)$$

where the pieces of the the cubic spline are defined as

$$\nu_j(x) = (x - k_j)_+^3 - \eta_j(x - k_{\min})_+^3 - (1 - \eta_j)(x - k_{\max})_+^3$$

Each of the k_j are the knots used to define the ranges where each cubic piece operates, the models are designed to be cubic between each of these knots and the η_j are defined to restrict the function to be linear outside of the range (below the first knot and above the final knot).

$$\eta_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}$$

We can also deviate from the proportional hazard requirement by defining the γ terms to be dependent on covariates; this is, of course, done via a linear model and demonstrates an extra layer of regression required to form such models:

$$\gamma_j(\mathbf{Z}) = \gamma_{j1} Z_1 + \gamma_{j2} Z_2 + ... + \gamma_{jm} Z_m$$

1.3.2 Competing Risks

A Competing Risk (CR) can be thought of simply as survival analysis where a cause of death, $D \in 1, ..., K$ is also observed [113]. When patients are recovering from a disease, more than one event can play a role, but often one event is of more interest than the other [113]. This competing event can also prevent the event-of-interest from occurring. For example, if we are modelling discharge from hospital after surgery (event-of-interest), patients can also die whilst in hospital (competing event) which prevents the former from happening.

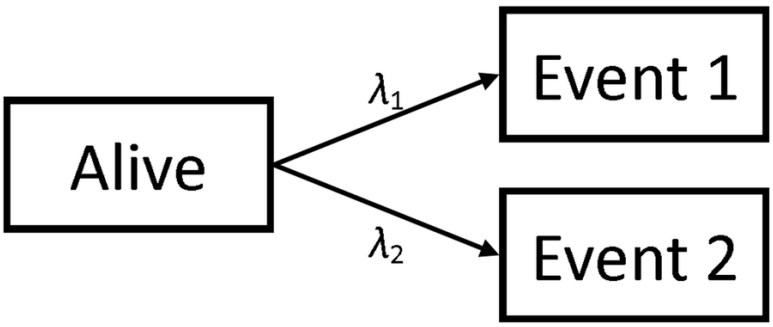

Figure 1.4: Examples of a simple Competing Risks Scenario

Depending on clinical context, non-administrative right censoring can be modelled as a competing risk [118] because censoring times are not always independent of event times [113]. If healthier patients are less likely to use medical services, then they are more likely to be lost to follow-up meaning a negative correlation with event time and vice versa if a less healthy person is more likely to leave a study (e.g. they become too ill to continue in the study).

If the competing event and event-of-interest are not independent, then this can cause bias in the Kaplan-Meier estimator [113]. Issues can arise with the naive Kaplan-Meier approach in CR, wherein the probability of having the events sum to more than 100%, even though the events can not occur together [113].

We can adjust the previous definition for a hazard function to become a cause-specific hazard function as:

$$\lambda_k(t) \lim_{\Delta t \to 0} \frac{\text{Prob}\,(T \leq t + \Delta T,\, D = k \,|\, T \geq t)}{\Delta t}$$

where k is the event we are assessing. Cause specific hazard (CSH) estimates may be found by treating the data as simple survival data and the competing events as a censoring event [118]. By adding all of the individual CSH, we get a total hazard across all competing events. If the events were all different causes of death, for example, this would be the total hazard of death, regardless of cause.

In order to translate these into an event-free Survival probability (or marginal survival probability), we perform the same calculation as above, using this total hazard

$$S(t) = \text{Prob}(T > t) = \exp\left(- \int_0^t \sum_{k=1}^{K} \lambda_k(u)\, \mathrm{d}u \right)$$

In traditional survival analysis, the probability of having the event, is simply $F(t) = 1 - S(t)$, however since failures can occur from different events, we now have to take this into account by utilising an event specific failure:

$$F_k(t) = \int_0^t S(u)\lambda_k(u)\,\mathrm{d}u$$

This is called the Cumulative Incidence Function (CIF). This can be broken down as the probability of surviving until time u, $S(u)$ and then the instantaneous probability of the event occurring at that time, $\lambda_k(u)$. We integrate this over all the possible event times on the range $(0, t]$. The CIF for a transition depends on all the other transition intensities through the $S(u)$ component of the integrand [118].

Ordinarily, a cdf, F, grows from 0 to 1 as the probability of having had an event increases over time. This assumes that eventually all patients will have the event (and thus $\lim_{t\to\infty} F(t) = 1$). However, when dealing with the CIF, patients who have one event are precluded from having any other event and therefore the assumption that all patients will have an event-of-interest does not hold. For this reason, the CIF is bounded within the range $[0, 1)$ and never reaches 1 and is thus known as a _sub_distribution [31].

Where we defined the hazard function earlier as the derivative of the log survival, we can define the *subdistribution* hazard similarly:

$$\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log\left(1 - F(t)\right) \qquad h_k(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log\left(1 - F_k(t)\right)$$

Fine & Gray [119] developed a method analogous to the Cox proportional hazards model for these subdistributions. The Proportional Subdistribution Hazards method (PSH) calculates the hazard ratios for events occurring based on the CIF, F_k [31], and subdistribution hazard where the Cox model uses the cumulative distribution function, F [31] and the cause-specific hazard.

$$\tilde{h}_k(t|\mathbf{Z}) = \tilde{h}_{k0}(t)\exp\left(\boldsymbol{\beta}_k^T\mathbf{Z}\right)$$

In this context, patients who have the competing event remain in the risk set for having the event-of-interest (and can be considered to be event-free "forever"), whereas in the CSH, they do not remain in the risk set.

An example of using a CR model in the real world involves mortality after an Acute Myocardial Infarction(AMI) [118]. Patients could either have a sudden Cardiovascular Disease (CVD) related death, a non-sudden CVD related death or a non-CVD related death (i.e. death from other causes). Using age and gender as covariates for a simple CSH Cox model [120], HRs can were calculated for each cause of death distinctly, and for an All-Cause hazard.

Having a more prognostic attitude towards medicine, as opposed to a diagnostic one, could help to reduce wasted expenditure under a CR scenario. For example, a patient with high blood pressure automatically has a lower risk of death from prostate cancer (because they are more likely to die from CVD) than a similar patient with normal blood pressure and so the first patient would gain less from a prostatectomy [33].

1.3.3 Multi-State Models

Just as CRs are an extension of survival analysis, Multi-State Models are an extension of CRs [113]. However, not all MSMs contain CRs. Figure 1.5 shows two examples of MSM systems, where the first models a progressive disease, which does not contain any CRs, and the second is known as an Illness-Death model and is one of the simplest MSMs.

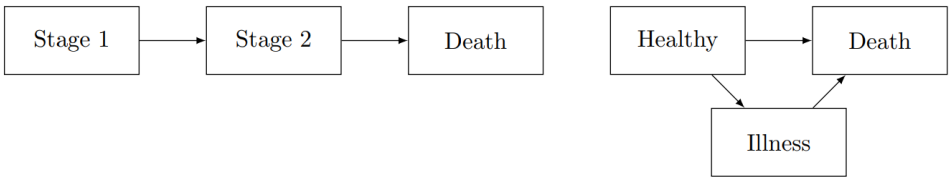

Figure 1.5: Two examples of Multi-State Models with three states.

An absorbing state is a state which has no transitions coming out of it (such as death or discharge), a state which is not absorbing is called transient [112]. States where a patient can begin are call initial states. Not all transient states are initial states, but all initial states must be are transient.

A CRs scenario is therefore a type of MSM where there is only a single initial state and multiple absorbing states which a patient can transition into [112], [118]. In a traditional survival analysis study, we may model patients moving to a different state as merely updating a covariate associated with that patient. However, this assumes that the underlying mechanism causing patients to move on is the same (e.g. in a Cox Model, the baseline hazard is the same). In a complex system, this is often not the case. MSMs can be used to model this scenario if we believe that the different states of the condition have an inconsistent effect on the outcome.

A multistate process is defined as a mapping $X : T \to K$ with $T = (0, \infty)$ is the time and a finite state space $K = \{1, ..., k\}$. The transition probabilities which we defined earlier, for transitioning to death from different causes, can now be extended to depend on the state that they are leaving, and the time we are estimating from.

$$P_{ij}(s,t) = \mathrm{P}\left(X(t) = j | X(s) = i,\ \mathcal{H}_t\right)$$

where $\mathcal{H}_t$ is the history of X for a given patient. We have to introduce the starting time, s in the above, as patients will not always enter a state at time $t = 0$. We could adopt this notation for CRs with:

$$F_k(t) = P_{1k}(0,t) \qquad\qquad P_{1k}(s,t) = \frac{F_k(t)}{F_k(s)}$$

This means we can also define the transition specific hazard function:

$$\lambda_{ij}(t) = \lim_{\Delta t \to 0} \frac{P_{ij}(t, t + \Delta t)}{\Delta t}$$

which has the same meaning as with CRs. It is the hazard of moving between states at time (i.e. from state i to j).

With this definition, we can give a mathematical definition of an absorbing state. $h \in K$ is an absorbing state iff $\lambda_{hj}(t) = 0 \quad \forall t \in (0, \infty), \; j \in K$. Even in transient states, some transition intensities will be 0 for all t, implying that a transition can never occur, such as in the right hand model of Figure 1.5. If all transitions in a model are one-way, then it is called a unidirectional model, if one or more can be reversed, then it is called a reversible MSM [112]. Figure 1.6 shows a reversible Illness-Death model, similar to the one in Figure 1.5, but patients can move from Illness to Healthy.

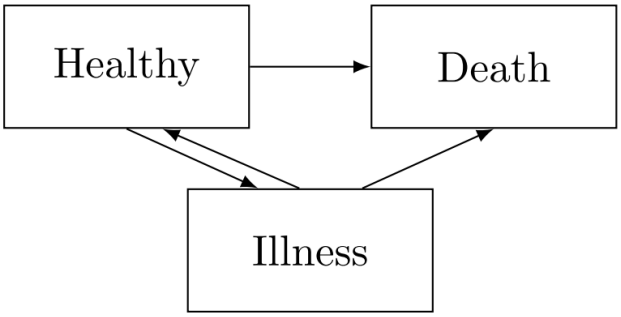

Figure 1.6: Example of an MSM with a reversible transition

Multi-state models can be split into two approaches. The non-homogeneous Markov approach assumes that $t = 0$ is when the patient enters the study and that time is tracked continuously across transitions. The Semi-Markov approach resets the time to $t = 0$ each time a new state is entered [55]. Clinical knowledge should be used to decide which method to use, and if a state is considered to be medically transformative (such as a transplant), then the Semi-Markov approach can be justified [121], however it may be useful to include a covariate to indicate how long a patient has been in other states. The Markov and Semi-Markov approaches are commonly referred to as "clock-forward" and "clock-reset" approaches respectively.

Although SLICC Damage Index (SDI) [122] could be considered to be a continuous or categorical variable, Bruce at al [37] used the fact that SDI is permanently increasing to model an MSM with the integer-valued SDI as an indicator of a patients state (grouping all patients with SDI > 5 into a single state for simplicity). see Figure 1.7 Since SDI is a representation of permanent damage, this model is unidirectional MSM. Their model used Cox regression to model a distinct hazard function for each of their transitions.

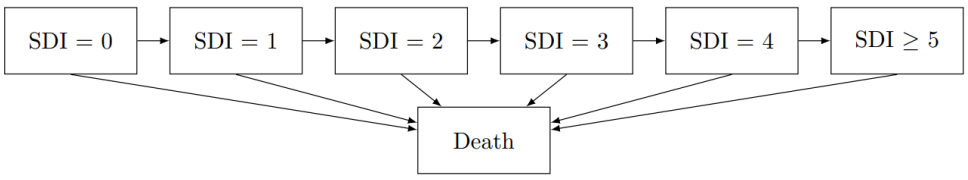

Figure 1.7: The progressive SDI model used by Bruce et al, 2015

It must be noted that it is not always possible to know the exact time of a transition leading to interval censoring (e.g. patient is in state A at t = 10 but is in state B at t =

20)[123]. This model developed by Bruce at al assumed that patients can transition from the state SDI $= n$ to the state SDI $= n + 1$,but there is no mention of patients being able to skip over states (e.g. straight from SDI $= 1$ to SDI $= 3$) which could potentially happen often where a patient is only assessed once a year. Without these extra transitions, there would need to be an assumption that patients who appear to skip out steps actually move through them between check ups which implies some sort of interval censoring[37].

Most MSM models can be broken down into a collection of Illness-Death models (with additional CRs depending on the shape of the model), which is why it is so common in examples [112]. In a non-reversible Illness-Death Model, we can define the transition probabilities depending on the three transition intensities, λ_{ij} [112], [113]. For these derivations, we will use the convention as in the matrix above, where Alive is state 1, Illness is state 2 and Death is state 3.

$$
\begin{aligned}
P_{11}(s,t) &= \exp\left(-\int_s^t \lambda_{12}(u) + \lambda_{13}(u)\ \mathrm{d}u\right) \\
P_{22}(s,t) &= \exp\left(-\int_s^t \lambda_{23}(u)\ \mathrm{d}u\right) \\
P_{12}(s,t) &= \int_s^t P_{11}(s,u)\lambda_{12}(u)P_{22}(u,t)\ \mathrm{d}u \\
P_{23}(s,t) &= \int_s^t P_{22}(s,u)\lambda_{23}(u)\ \mathrm{d}u \\
P_{13}^1(s,t) &= \int_s^t P_{11}(s,u)\lambda_{13}(u)\ \mathrm{d}u \\
P_{13}^2(s,t) &= \int_s^t P_{12}(s,u)P_{23}(u,t)\ \mathrm{d}u \\
P_{13}(s,t) &= P_{13}^1(s,t) + P_{13}^2(s,t)
\end{aligned}
$$

Notice that in the model, there are two distinct pathways that a patient can travel to move from Alive to Death. We can track these mathematically, where P_{13}^1 estimates patients going directly to Death, and P_{13}^2 estimates patients going via Illness. We can then combine them into a single probability.

Similar to an earlier example, in P_{12}, we have the probability of staying in the Healthy state from s to u, multiplied by the hazard of transitioning at time u and then by the probability of staying in the Illness state, from u to t integrated over all of u between s and t, this is quite an intuitive definition.

If we include reversibility in an Illness-Death model, we need to take this into account as a Competing Risk when a patient is in the Illness state. We would need to include the probability of moving from Illness to Healthy and adjust the definition of staying in the Illness state. However, the above formula only includes the probability of a patient *remaining* in a state between given time points, and does not account for patients *returning* to that state between time s and t. This leads to an iterative definition, where n is the number of times a

patient has left the former state:

$$
\begin{aligned}
P_{11}^{0}(s,t) &= \exp\left(-\int_s^t \lambda_{12}(u) + \lambda_{13}(u)\ \mathrm{d}u\right) \\
P_{22}^{0}(s,t) &= \exp\left(-\int_s^t \lambda_{23}(u) + \lambda_{32}(u)\ \mathrm{d}u\right) \\
P_{12}^{0}(s,t) &= \int_s^t P_{11}^{0}(s,u)\lambda_{12}(u)P_{22}^{0}(u,t)\ \mathrm{d}u \\
P_{21}^{0}(s,t) &= \int_s^t P_{22}^{0}(s,u)\lambda_{21}(u)P_{11}^{0}(u,t)\ \mathrm{d}u \\
P_{12}^{n}(s,t) &= \int_s^t P_{11}^{n}(s,u)\lambda_{12}(u)P_{22}^{0}(u,t)\ \mathrm{d}u \\
P_{21}^{n}(s,t) &= \int_s^t P_{22}^{n}(s,u)\lambda_{21}(u)P_{11}^{0}(u,t)\ \mathrm{d}u \\
P_{11}^{n}(s,t) &= \int_s^t P_{12}^{n-1}(s,u)P_{21}^{1}(u,t)\ \mathrm{d}u \\
P_{22}^{n}(s,t) &= \int_s^t P_{21}^{n-1}(s,u)P_{12}^{1}(u,t)\ \mathrm{d}u \\
P_{11}(s,t) &= \sum_{n=1}^{\infty} P_{11}^{n}(s,t) \\
P_{22}(s,t) &= \sum_{n=1}^{\infty} P_{22}^{n}(s,t) \\
P_{12}(s,t) &= \sum_{n=1}^{\infty} P_{12}^{n}(s,t) \\
P_{21}(s,t) &= \sum_{n=1}^{\infty} P_{21}^{n}(s,t)
\end{aligned}
$$

For convenience, the superscripts on the right of the equations *always* add to the superscripts on the left. This iterative definition and infinite summation is why reversible MSMs become much more complicated than a simple mono-directional MSM.

Another common model involves having two conditions, A & B, which may or may not be independent [112]. A simple way to model these two stages is as four states, shown on the left of Figure 1.8, where once both conditions are established, they are in State "A & B." If, however, the order of developing A and B is important, then the state "A & B" can be split into two states: "A then B" and "B then A," as in the right model of 1.8

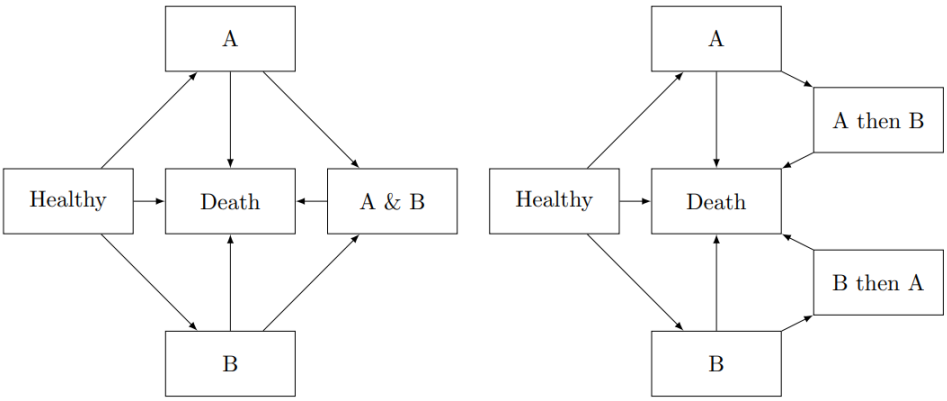

Figure 1.8: Two constructions of an A & B Model

Liquet et al [55] developed a Multi-State frailty model of patients developing a Ventilator-Associated Pneumonia infection (VAP), being discharged or dying. They found that there was significant heterogeneity in the development of VAP and in discharge, but not significant heterogeneity in dying with or without VAP and discharge with VAP. This demonstrates that clustering can occur in some transitions but not in others and provides evidence that frailty should also be included in a large scale study. This study concluded that VAP is frequent in Intensive Care Units (ICUs) and is associated with an increase in ICU mortality, length of stay and cost [55].

Anwar and Mahmoud [111] developed a stochastic model which treated CKD progression as a series of states in an MSM with three state being dependent on the Glomerular Filtration Rate (GFR) of the kidneys (mild, moderate and severe reduction) and End-Stage Renal Disease (ESRD) being the fourth stage and death being the final, absorbing state. They used this model to estimate survival probability, see Figure 1.9. Their model combined the hazards from the "non-dead" states into a larger stochastic hazard to predict patient transitions from Alive to Death.

CRs and MSMs are versatile mechanics which can be used to great advantage of researchers if done properly. By mapping the journey through the system of every patient in a population, detailed predictions can be made for future patients. By using modelling methods such as the R-P method which provides absolute probabilities, clinicians can provide patients with the probability of them having a particular event, as well as the probability of them being in a given state at a certain time in the future. MSMs can become very complicated very quickly, and even something as simple as adding reversibility to a transition increases the convolutions that can occur.

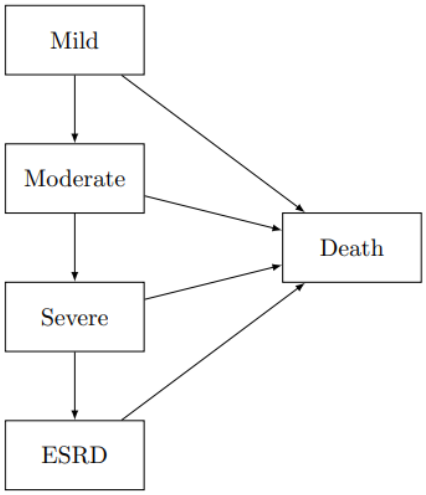

Figure 1.9: The progressive CKD model used by Anwar et al, 2014

1.4 Chronic Kidney Disease

Chronic diseases, especially non-communicable ones, have now become the major cause of morbidity and mortality around the world [124]. In particular, Chronic Kidney Disease (CKD) is a global health concern [16] and is thus a major burden on healthcare utilisation worldwide [125]. This is unsurprising given that, in the UK, 7,411 patients commenced RRT in 2004 alone which equates to a rate of 115 per million people [126]. Part of this prevalence is believed to be increasing due to increased incidences of diabetes [16] which contribute 26.9% of new RRT diagnosis in the UK in 2014 [126]. In 2013, the NHS spent 2% of its budget on kidney replacement therapy [16], [127] and in 2008, 5.9% of the Medicare expenditure was spent on managing patients with End-Stage Renal Disease (ESRD) [16], [128]. The progression of CKD amongst sufferers is believed to be homogeneous with respect to time [111], meaning that it increases continuously at steady rate.

CKD treatment typically consists of either palliative care or a type of RRT. In the real world, it is difficult to make decisions on RRT for patients suffering from ESRD since, as with any disease, there is a lot of variability in the individuals [40]. This variability is particularly prominent amongst older patients, which leads to variation in treatment methods from different physicians [129]. Because of this, it is important to identify, as early as possible, patients who are likely to progress from CKD to ESRD [16]. Tamura et al [40] provides a framework for deciding which RRT patients should receive based on three factors: life expectancy, risks and benefits of competing treatment strategies and patient preference. This framework does not require precision, but rather a general idea of whether a patient is above or below average (median). These three factors allows for three key choices to be made for the patient: choice of dialysis modality (i.e. HD vs PD), choice of vascular access for HD, and whether or not to be referred for kidney transplantation.

Transferring from one dialysis modality to the other initially increases burden on patients and, for the first few weeks, has a higher mortality rate [130]. Before beginning HD, patients

and physicians must decide on the vascular access method which is basically how the HD will be administered. There are three main methods of vascular access: CVCs, AVFs and AVGs [40]. In the US, 80% of patients who are given HD, begin it with a CVC [131]. CVCs are usually used as a temporary placement until a more permanent fistula or graft can be given to the patient [40]. However, it can take time for AVF and AVG patency to occur and so their effects are not immediate. Current guidelines recommend using AVFs over AVGs as the method of permanent access which are both preferred over the temporary access provided by CVC, unless HD is predicted to be only a short-term treatment (i.e. because of expected kidney transplant or extremely high expected mortality) [132]. It is clear that these mortality estimates of patients are currently wildly incorrect as it has been found that two-thirds of deceased patients who had undergone AVF placement had died before it was even used [133].

It is suggested that PD gives an early benefit over HD using CVC due to the high infection rates caused by CVC. However, this benefit might be balanced out by the higher risk of modality failure and a common need to transfer to HD later, which merely pushes the higher CVC risk back [134]. In recent years, It has been observed that survival amongst patients given PD has increased to levels similar to HD [135], although this is likely biased due to the difference in patient's selected for the each modality [40].

Kidney Transplants are often hard to come by as there can be difficulties in finding compatible donors [136]. Living donors provide a better prognosis for recipients than deceased ones, but even deceased-donor transplantation implies a 48-82% decrease in mortality compared to remaining on dialysis [137], [138]. For each patient, donors can be classified as being from a Standard Criteria Donor (SCD) or an Expanded Criteria Donor (ECD) list [139]. An ECD is, as the name implies, a much broader list of patients than appear on an SCD. Using an ECD comes with a shorter time on the waiting list for a transplant, but a higher risk of allograft loss and so a decision must be made about a patient of whether they are at higher risk of mortality if they remain on the waiting list for a longer period of time, or whether the risk of an unsuccessful transplant is worth it [40]. As with transferring between dialysis modalities, there is an extremely high increase in risk for the first two weeks after transplantation (compared with staying on dialysis), this risk reduces until 7-8 months after transplantation, where the cumulative mortality of both options becomes equivalent, and afterwards is lower for the transplanted patients [40]. It is worth noting that there is no upper age limit on kidney transplantation [40] and it has actually been found that kidney transplantation was cost-effective amongst patients over 6530. This makes sense as, for patient's over 65, the average time spent on the waiting list for a new kidney is 7-8 months [40].

In the UK in 2014, 71.8% of RRT patients had begun with HD, 20.0% were given PD, 8.2% were lined up to receive a kidney transplant [126]. Of the patients who were initially assigned to received HD in 2009, 54.4% had died by 2014 and 34.4% of those still alive had been transferred to a different modality, PD had a lower mortality rate, 35.1%, but a higher transfer rate, 75.3% [126]. Although these numbers do not account for the differences between the patients these two modalities were given to, it shows that there are major differences between modalities and that transitioning between treatments is common. These differences between modalities and the prevalence of transitions line up quite well with the idea of using an MSM as a representation

for this process.

1.5 Conclusion

This chapter has introduced the key concepts that will be built upon through this thesis. The main three topics that have been discussed are Clinical Prediction Models (CPMs), Multi-State Models (MSMs) and Chronic Kidney Disease (CKD).

It is clear from this work that CPMs require many steps to justify and that researchers need to focus more on the validation of existing models when they are available, and less on the development of novel ones. This, of course, is a problem that perpetuates within many of the modern scientific fields [98].

MSMs provide a robust and mathematical way to analyse movement through complicated clinical pathways. They improve upon the methods developed within survival analysis by including the addition of transient states and multiple end-points. This can allow researchers to visualise a more in-depth perspective on patients journeys through their illnesses.

The application of MSMs to CPMs is not common with the literature as it currently stands. While each of these two fields have been studied in depth in their own right, the overlap and intersection remains relatively desolate. It is clear that there are many gaps in this area of literature and this thesis aims to address several of them.

As discussed in this chapter, when creating a clinical prediction model, there are various stages that a model must go through before it is useable to demonstrate it's clinical utility. Two key steps in this process are the development and validation step. While there is a wealth of advice on developing and validating clinical prediction models in general, as well as more specific work in particular models, there is very little on how to develop and validate a multi-state clinical prediction model.

For example, there are currently no established and standardised methods for assessing the quality of an MSCPM at the point of validation; whereas there are multiple metrics to establish many measurements for other specific CPM techniques. This idea is explored further in Chapters 3 and 4, wherein our aim is to build from commonly used metrics in traditional survival analysis based models to design collapsible extensions.

Further to this, the specific application of multi-state models to chronic kidney disease is not common, and the currently existing models were developed with sub-optimal techniques such as the use of multinomial analysis rather than a structured multi-state model (such as the model developed by Grams et al [140]) or the use of arbitrary cut-off points to turn a continuous variable into a series of states (such as the model developed by Begun et al [141]). These models are discussed further in Chapter 5, where we develop our own MSCPM, which utilises much more optimal techniques for MSCPM development and validation.

Blank Page

Chapter 2

How Unmeasured Confounding in a Competing Risks Setting Can Affect Treatment Effect Estimates in Observational Studies

MA Barrowman, N Peek, M Lambie, GP Martin, M Sperrin

Published as: **MA Barrowman**, N Peek, M Lambie et al, How Unmeasured Confounding in a Competing Risks Setting Can Affect Treatment Effect Estimates in Observational Studies, BMC Medical Research Methodology (2019) doi: 10.1186/s12874-019-0808-7

Abstract

Background

Analysis of competing risks is commonly achieved through a cause specific or a subdistribution framework using Cox or Fine & Gray models, respectively. The estimation of treatment effects in observational data is prone to unmeasured confounding which causes bias. There has been limited research into such biases in a competing risks framework.

Methods

We designed simulations to examine bias in the estimated treatment effect under Cox and Fine & Gray models with unmeasured confounding present. We varied the strength of the unmeasured confounding (i.e. the unmeasured variable's effect on the probability of treatment and both outcome events) in different scenarios.

Results

In both the Cox and Fine & Gray models, correlation between the unmeasured confounder and the probability of treatment created biases in the same direction (upward/downward) as the effect of the unmeasured confounder on the event-of-interest. The association between correlation and bias is reversed if the unmeasured confounder affects the competing event. These effects are reversed for the bias on the treatment effect of the competing event and are amplified when there are uneven treatment arms.

Conclusion

The effect of unmeasured confounding on an event-of-interest or a competing event should not be overlooked in observational studies as strong correlations can lead to bias in treatment effect estimates and therefore cause inaccurate results to lead to false conclusions. This is true for cause specific perspective, but more so for a subdistribution perspective. This can have ramifications if real-world treatment decisions rely on conclusions from these biased results. Graphical visualisation to aid in understanding the systems involved and potential confounders/events leading to sensitivity analyses that assumes unmeasured confounders exists should be performed to assess the robustness of results.

Supplementary Material

Supplementary Material is available in Appendices A & B.

2.1 Background

Well-designed observation studies permit researchers to assess treatment effects when randomisation is not feasible. This may be due to cost, suspected non-equipoise treatments or any number of other reasons [142]. While observational studies minimise these issues by being cheaper to run and avoiding randomisation (which, although unknown at the time, may prescribe patients to worse treatments), they are potentially subject to issues such as unmeasured confounding and increased possibility of competing risks (where multiple clinically relevant events occur). Although these issues can arise in any study, Randomised Controlled Trials (RCTs) attempt to mitigate these effects by using randomisation of treatment and strict inclusion/exclusion criteria. However, the estimated treatment effects from RCTs are of potentially limited generalisability, accessibility and implementability [143].

A confounder is a variable that is a common cause of both treatment and outcome. For example, a patient with a high Body Mass Index (BMI) is more likely to be prescribed statins [5], but are also more likely to suffer a cardiovascular event. These treatment decisions can be affected by variables that are not routinely collected (such as childhood socio-economic status or the severity of a comorbidity [144]. Therefore, if these variables are omitted form (or unavailable for) the analysis of treatment effects in observational studies, then they can bias inferences [145]. As well as having a direct effect on the event-of-interest, confounders (along with other covariates) can also have further reaching effects on a patient's health by

changing the chances of having a competing event. Patients who are more likely to have a competing event are less likely to have an event-of-interest, which can affect inferences from studies ignoring the competing event. In the above BMI example, a high BMI can also increase a patient's likelihood of developing (and thus dying from) cancer [146].

The issue of confounding in observational studies has been researched previously [147]–[148], where it has been consistently shown that unmeasured confounding is likely to occur within these natural datasets and that there is poor reporting of this, even after the introduction of the The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Guidelines [149], [150]. Hence, it is widely recognised that sensitivity analyses are vital within the observational setting [151]. However these previous studies do not extend this work into a competing risk setting, meaning research in this space is lacking [152], particularly where the presence of a competing event can affect the rate of occurrence of the event-of-interest. These issues will commonly occur in elderly and comorbid patients where treatment decisions are more complex. As the elderly population grows, the clinical community needs to understand the optimal way to treat patients with complex conditions; here, causal relationships between treatment and outcome need to account for competing events appropriately.

The most common way of analysing data that contains competing events is using a cause specific perspective, as in the Cox methodology [153], where competing events are considered as censoring events and analysis focuses solely on the event-of-interest. The alternative is to assume a subdistributional perspective, as in the Fine & Gray methodology [119], where patients who have competing events remain in the risk set forever.

The aim of this paper is to study the bias induced by the presence of unmeasured confounding on treatment effect estimates in the competing risks framework. We investigated how unmeasured confounding affects the apparent effect of treatment under the Fine & Gray and the Cox methodologies and how these estimates differ from their true value. To accomplish this, we used simulations to generate synthetic time-to-event-data and then model under both perspectives. Both the Cox and Fine & Gray models provide hazard ratios to describe the effects of a covariate. A binary covariate will represent a treatment and the coefficients found by the model will be the estimate of interest.

2.2 Methods

We considered a simulation scenario in which our population can experience two events; one of which is the event-of-interest (Event 1), the other is a competing event (Event 2). We model a single unmeasured confounding covariate, $U \sim N(0, 1)$ and a binary treatment indicator, Z. We varied how much U and Z affect the probability distribution of the two events as well as how they are correlated. For example, Z could represent whether a patient is prescribed statins, U could be their BMI, the event-of-interest could be cardiovascular disease related mortality and a competing event could be cancer-related mortality. We followed best practice for conducting and reporting simulations studies [154].

The data-generating mechanism defined two cause-specific hazard functions (one for each event), where the baseline hazard for event 1 was k times that of event 2, see Fig. 2.1. We

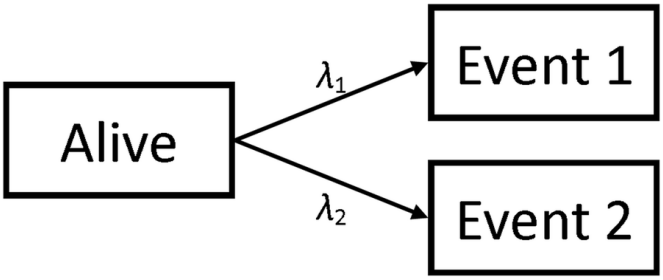

Figure 2.1: Transition State Diagram showing potential patient pathways

assumed a baseline hazard that was either constant (exponential distributed failure times), linearly increasing (Weibull distributed failure times) or biologically plausible [155]. The hazards used were thus:

$$\lambda_1(t|U,Z) = ke^{\beta_1 U + \gamma_1 Z}\lambda_0(t) \tag{2.1}$$

$$\lambda_2(t|U,Z) = ke^{\beta_2 U + \gamma_2 Z}\lambda_0(t) \tag{2.2}$$

$$\lambda_0(t) \begin{cases} 1 & \text{Exponential} \\ 2t & \text{Weibull} \\ \exp\left(-18 + 7.3t - 11.5t^{0.5}\log(t) + 9.5t^{0.5}\right) & \text{Plausible} \end{cases} \tag{2.3}$$

In the above equations, β and γ are the effects of the confounding covariate and the treatment effect respectively with the subscripts representing which event they are affecting. These two hazard functions entirely describe how a population will behave [156].

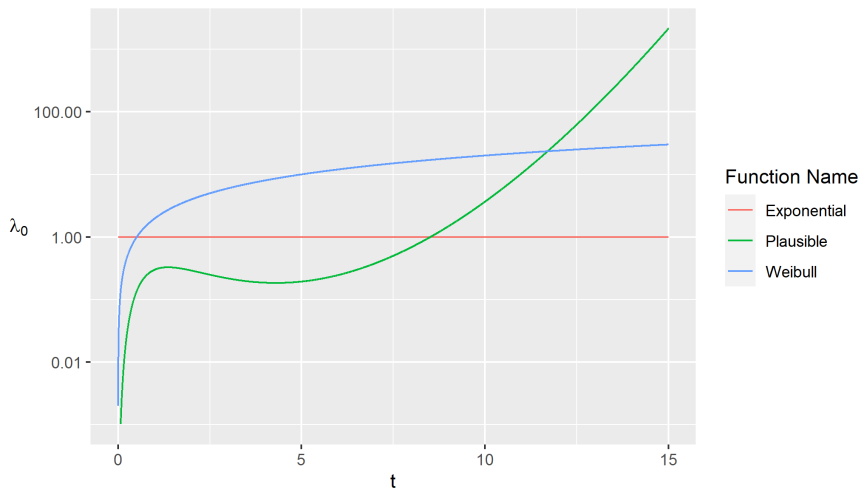

Figure 2.2: Plot of the baseline hazards used as part of this simulation study

We simulated populations of 10,000 patients to ensure small confidence intervals around our treatment effect estimates in each simulation. Each simulated population had a distinct value for β and γ. In order to simulate the confounding of U and Z, we generated these values

such that $\text{Corr}(U, Z) = \rho$ and $\text{P}(Z = 1) = \pi$ [157]. Population end times and type of event were generated using the relevant hazard functions. The full process for the simulations can be found in Appendix B. Due to the methods used to generate the populations, the possible values for ρ are bounded by the choice of π such that when $\pi = 0.5$, $|\rho| <= 0.797$ and when $\pi = 0.1$ (or $\pi = 0.9$), $|\rho| <= 0.57$. The relationship between the parameters can be seen in the Directed Acyclic Graph (DAG) shown in Fig. 2.3, where T is the event time and δ is the event type indicator (1 for event-of-interest and 2 for competing event). From this, we also

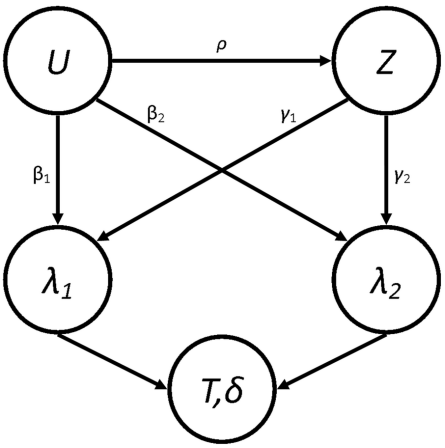

Figure 2.3: Directed Acyclic Graph showing the relationship between some of the parameters

explicitly calculated what we would expect the true subdistribution treatment effects, Γ_1 and Γ_2, to be in these conditions [158] (See Appendix A). It's worth noting that the values of Γ will depend on the current value of ρ since they are calculated using the expected distribution of end-times. However, it has been shown [156], [159] that, due to the relationship between the Cause-Specific Hazard (CSH) and the Subdistribution Hazard (SH), only one proportional hazards assumption can be true. Therefore the "true" values of the Γ will be misspecified and represent a least false parameter (which itself is an estimate of the time-dependent truth) [158].

We used the simulated data to estimate the treatment effects under the Cox and Fine & Gray regression methods. We specify that U is unmeasured and so it wasn't included in the analysis models. As discussed earlier, the Cox model defines the risk set at time t to be all patients who have not had any event by time t, whereas the Fine & Gray defines it to be those who have not had the event-of-interest (or competing event) by time t.

For our models, for the events, $i = 1, 2$, we therefore defined the CSH function estimate, $\hat{\lambda}_i$, and the SH function estimate, $\hat{h}_i$, to be

$$\hat{\lambda}_i(t|Z) = \hat{\lambda}_{i0}(t)e^{\hat{\gamma}_i Z} \qquad \hat{h}_i(t|Z) = \hat{h}_{i0}(t)e^{\hat{\Gamma}_i Z}$$

Where $\hat{\lambda}_{i0}(t)$ and $\hat{h}_{i0}(t)$ are the baseline hazard and baseline subdistribution hazard function estimates for the entire population (i.e. no stratification), and $\hat{\gamma}_i$ and $\hat{\Gamma}_i$ are the estimated treatment effects. From these estimates, we also extracted the estimate of the subdistribution treatment effect in a hypothetical RCT, where $\rho = 0$ and $\pi = 0.5$ to give $\hat{\Gamma}_{10}$ and $\hat{\Gamma}_{20}$. To investigate how the correlation between U and Z affects the treatment effect estimate, we

compared the explicitly prescribed or calculated values with the simulated estimates. Three performance measures for both events, along with appropriate 95% confidence intervals, were calculated for each set of parameters:

- $\theta_{\text{RCT},i} = \text{E}\left[\hat{\Gamma}_i - \hat{\Gamma}_{i0}\right]$ ˜ The average difference between the SH treatment effect estimate from an idealised, hypothetical RCT situation.
- $\theta_{\text{Exp},i} = \text{E}\left[\hat{\Gamma}_i - \Gamma_i\right]$ ˜ The average bias of the SH treatment effect estimate from the explicitly calculated value.
- $\theta_{\text{CSH},i} = \text{E}\left[\hat{\gamma}_i - \gamma_i\right]$ ˜ The average bias of the CSH treatment effect estimate from the predefined treatment effect.

As mentioned above, the value of Γ will depend on the current value of ρ and so the estimation of the explicit bias will be a measure of the total bias induced on our estimate of the subdistribution treatment effect in those specific set of parameters. We also evaluate the bias compared to an idealised RCT to see how much of this bias could be mitigated if we were to perform an RCT to assess the effectiveness of the hypothetical treatment. Finally, we found the explicit bias in the cause specific treatment effect to again see the total bias applied to this measure. We did not compared the CSH bias to an idealised RCT as we believed that this could easily be inferred from the CSH explicit results, whereas this information wouldn't be as obvious in the SH treatment effect due to the existence of a relationship between Γ and ρ.

Eight Scenarios were simulated based on real-world situations. In each scenario, ρ varied across 5 different values ranging from 0 to their maximum possible value (0.797 for all Scenarios apart from Scenario 5, where it is 0.57, due to the bounds imposed by the values of π). One other parameter (different for different scenarios) varied across 3 different values, and all other parameters were fixed as detailed in Table 1. Each simulation was run 100 times and the performance measures were each pooled to provide small confidence intervals. This gives a total of 1,500 simulations for each of the 8 scenarios. Descriptions of the different scenarios are given below:

1. No Effect. To investigate whether treatment with no true effect ($\gamma_1 = \gamma_2 = 0$) can have an "artificial" treatment effect induced on them in the analysis models through the confounding effect on the event-of-interest. β_1 varied between -1, 0 and 1.
2. Positive Effect. To investigate whether treatment effects can be reversed when the treatment is beneficial for both the event-of-interest and the competing event ($\gamma_1 = \gamma_2 = -1$). β_1 varied between -1, 0 and 1.
3. Differential Effect. To investigate how treatment effect estimates react when the effect is different for the event-of-interest ($\gamma_1 = -1$) and the competing event ($\gamma_2 = 1$). β_1 varied between -1, 0 and 1.
4. Competing confounder. To investigate whether treatments with no true effect ($\gamma_1 = \gamma_2 = 0$) can have an "artificial" treatment effect induced on them by the effect of a confounded variable on the competing event only ($\beta_1 = 0$). β_2 varied between -1, 0 and 1.

5. Uneven Arms. To investigate how having uneven arms on a treatment in the population can have an effect on the treatment effect estimate ($\gamma_1 = -1$, $\gamma_2 = 0$). π varied between $^1/_{10}$, $^1/_2$ and $^9/_{10}$.

6. Uneven Events. To investigate how events with different frequencies can induce a bias on the treatment effect, despite no treatment effect being present ($\gamma_1 = \gamma_2 = 0$). k varied between $^1/_2$, $^1/_2$ and 2.

7. Weibull Distribution. To investigate whether a linearly increasing baseline hazard function affects the results found in Scenario 1. β_1 varied between -1, 0 and 1.

8. Plausible Distribution. To investigate whether a biologically plausible baseline hazard function affects the results found in Scenario 1. β_1 varied between -1, 0 and 1.

Table 2.1: Details of parameters for each Scenario

Sc	ρ					Baseline	γ_1	γ_2	β_1			β_2			π			k		
1	0.00	0.20	0.40	0.60	0.80	Constant	0	0	-1	0	1	0			$^1/_2$			1		
2	0.00	0.20	0.40	0.60	0.80	Constant	-1	-1	-1	0	1	0			$^1/_2$			1		
3	0.00	0.20	0.40	0.60	0.80	Constant	-1	1	1	0	1	0			$^1/_2$			1		
4	0.00	0.20	0.40	0.60	0.80	Constant	0	0	0			1	0	1	$^1/_2$			1		
5	0.00	0.14	0.29	0.42	0.57	Constant	0	0	1			0			$^1/_{10}$	$^1/_2$	$^9/_{10}$	1		
6	0.00	0.20	0.40	0.60	0.80	Constant	0	0	1			0			$^1/_2$			$^1/_2$	1	2
7	0.00	0.20	0.40	0.60	0.80	Weibull	0	0	-1	0	1	0			$^1/_2$			1		
8	0.00	0.20	0.40	0.60	0.80	Plausible	0	0	-1	0	1	0			$^1/_2$			1		

2.3 Results

The first row of Fig. 3 shows the results for Scenario 1 (No Effect). When $\beta_1 = \beta_2 = 0$ (the green line), correlation between U and Z doesn't imbue any bias on the treatment effect estimate for either event under any of the three measures, since all of the subdistribution treatment effects (estimated, calculated and hypothetical RCT) are approximately zero. When $\beta_1 > 0$, there is a strong positive association between correlation (ρ) and the RCT and CSH biases for the event-of-interest and a negative association for the RCT bias for the competing event. Similarly, these associations are reversed when $\beta_1 < 0$. There was no effect on θ_{CSH} for the competing event in this Scenario regardless of ρ or β_1. These results are similar to those found in Scenario 2 (Positive Effect) and Scenario 3 (Negative Effect) shown in Figs. 4 and 5. However, in both of these Scenarios, there is an overall positive shift in θ_{CSH} when $\beta_1 \neq 0$.

The magnitude of θ_{Exp} is greatly reduced and is the reverse of the other associations when $\beta_1 \neq 0$ in Scenario 1 for the event-of-interest and when $\beta_1 > 0$ it stays extremely small for low values of ρ, and becomes negative for large ρ for the competing event. In Scenario 2, θ_{Exp} behaves similarly to Scenario 1 for both events when $\beta_1 < 0$ and the event-of-interest, but for the competing event, when $\beta_1 > 0$, the θ_{Exp} is much tighter to 0. The competing event data for θ_{Exp} in Scenario 3 is similar to Scenario 2 with $\beta_1 > 0$ shifted downwards, but the event-of-interest has a near constant level of bias regardless of ρ, apart from in the case when $\beta_1 < 0$, the bias switches direction. In Scenario 4 (Competing confounder), as would be expected, the

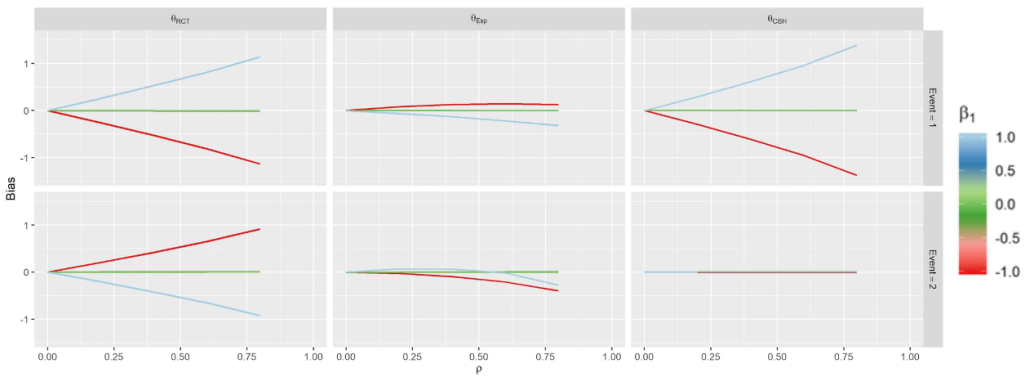

Figure 2.4: How changes in ρ affect bias in Scenario 1 - No Effect

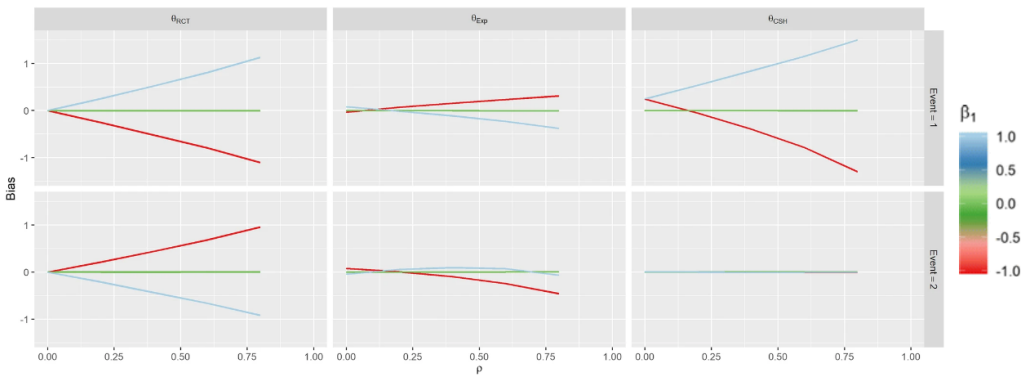

Figure 2.5: How changes in ρ affect bias in Scenario 2 - Positive Effect

results for the event-of-interest and the results for the competing event are swapped from those of Scenario 1 as shown in Fig. 6. Scenario 5 (Uneven Arms) portrays a bias similar to Scenario 1 where $\beta_1 = 1$, however, the magnitude of the RCT and CSH bias is increased when $\pi \neq 0.5$ as shown in Fig. 7.

The parameters for Scenario 6 (Uneven Events) were similar to the parameters for Scenario 1 (No Effect), when $\beta_1 = 1$. This also reflects in the results in Fig. 8 which look similar to the results for this set of parameters in Scenario 1. This bias is largely unaffected by the value of k. The results of Scenario 7 (Weibull Distribution) and Scenario 8 (Plausible Distribution) were nearly identical to those of Scenario 1 as shown in Figs. 9 and 10. As per our original hypotheses, Scenario 1 demonstrated that it is possible to induce a treatment effect when one isn't present through confounding effects on all biases, apart from the competing event CSH. In Scenario 2, with high enough correlation, the CSH event-of-interest bias could be greater than 1, meaning that the raw CSH treatment effect was close to 0, despite an actual treatment effect of -1, similarly large positive biases in the SH imply a treatment with no benefit and/or detrimental effect, despite the true treatment being beneficial for both events. This finding is similar for Scenario 3 with large biases changing the direction of the treatment effect (beneficial vs detrimental).

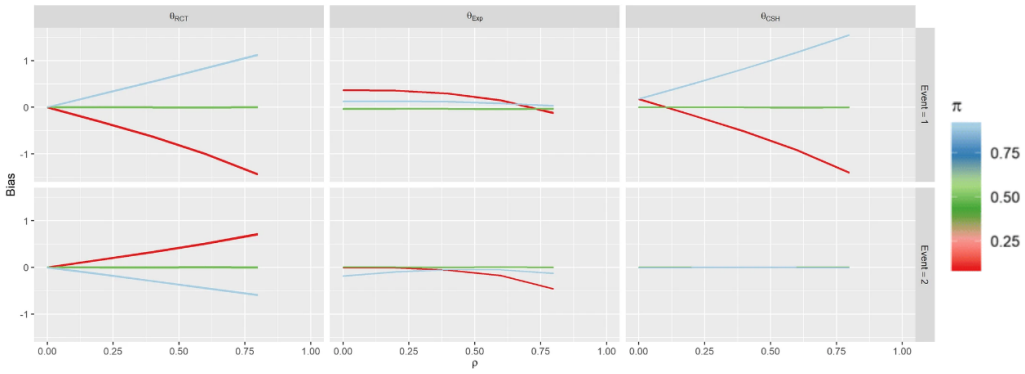

Figure 2.6: How changes in ρ affect bias in Scenario 3 - Differential Effect

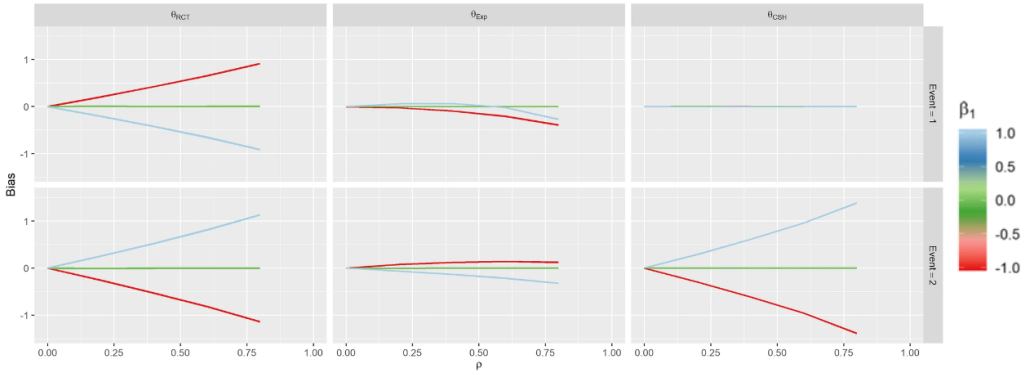

Figure 2.7: How changes in ρ affect bias in Scenario 4 - Competing confounder

Scenario 4 demonstrated that even without a treatment effect and with no confounding effect on the event-of-interest, a treatment effect can be induced on the SH methodology, which can imply a beneficial/detrimental treatment, depending on whether the confounder was detrimental/beneficial. Fortunately, it does not induce an effect on the CSH treatment effect for the event-of-interest. Scenarios 5 and 6 investigated other population level effects; differences in the size of the treatment arms and differences in the magnitude of the hazards of the events. Scenario 5 demonstrated that having uneven treatment arms can exacerbate the bias induced on both the θ_{RCT} and θ_{CSH} for both events and Scenario 6 showed that the different baseline hazards had little effect on the levels of bias in the results. This finding was supported by the additional findings of Scenarios 7 and 8, which showed that the underlying hazard functions did not affect the treatment effect biases compared to a constant hazard.

2.4 Discussion

This is the first paper to investigate the issue of unmeasured confounding on a treatment effect in a competing risks scenario. Herein, we have demonstrated that regardless of the actual effect of a treatment on a population that is susceptible to competing risks, bias can be induced by

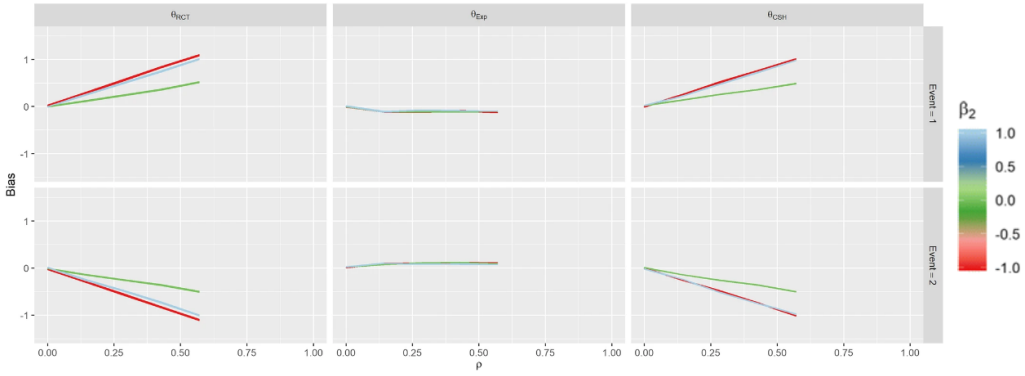

Figure 2.8: How changes in ρ affect bias in Scenario 5 - Uneven Arms

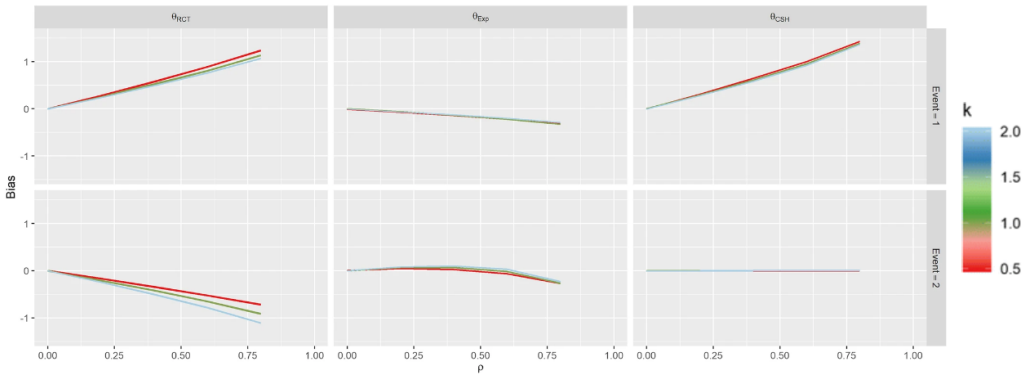

Figure 2.9: How changes in ρ affect bias in Scenario 6 - Uneven Events

the presence of unmeasured confounding. This bias is largely determined by the strength of the confounding relationship with the treatment decision and size of confounding effect on both the event-of-interest and any competing events. This effect is present regardless of any difference in event rates between the events being investigated and is also exacerbated by misbalances in the number of patients who received treatment and the number of patients who did not.

Our study has shown how different the case would be if a similar population (without inclusion/exclusion criteria) were put through an RCT and how the correlation between an unmeasured confounder and the treatment is removed, as would be the case in a pragmatic RCT. By combining the biases from an RCT and the explicitly calculated treatment effect, we can also use these results to infer how much of the bias found here is from omitted variable bias [160] and how much is explicitly due to the correlation between the covariates. Omitted variable bias occurs when a missing covariate has an effect on the outcome, but is not correlated with the treatment (and so is not a true confounder). It can occur even if the omitted variable is initially evenly distributed between the two treatment arms because, as patients on one arm have events earlier than the other, the distributions of the omitted variable drift apart. This makes up some of the bias caused by unmeasured confounding, but not all of it. For example, in Scenario 3 (Differential Effect), the treatment lowered the hazard of the event-of-interest,

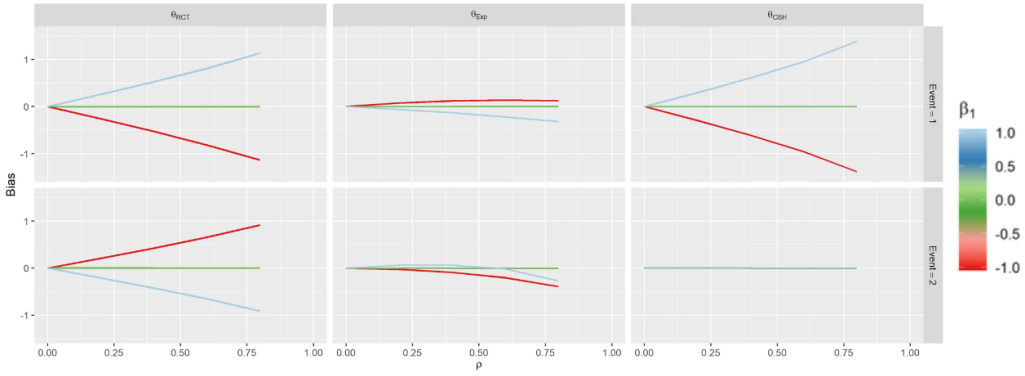

Figure 2.10: How changes in ρ affect bias in 7 - Weibull Distribution

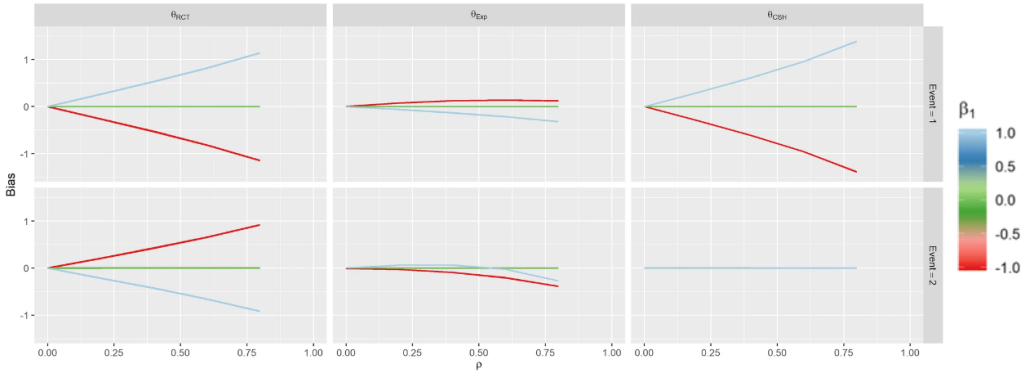

Figure 2.11: How changes in ρ affect bias in Scenario 8 - Plausible Distribution

but increased the hazard of the competing event; with a median level of correlation ($\rho = 0.4$), the event-of-interest bias from the RCT when there is a negative confounding effect ($\beta_1 < 0$) is -0.628 and the bias from the explicit estimate is 0.295 and therefore, the amount of bias due purely to the correlation between the unmeasured confounder and the treatment is actually -0.923. In this instance, some of the omitted variable bias is actually mitigating the bias from the correlation; if we have two biasing effects that can potentially cancel each other out, we could encounter a Type III error [161] which is very difficult to prove and can cause huge problems for reproducibility (if you eliminate a single source of bias, your results will be farther from the truth).

Our simulations indicate that a higher (lower) value of β_1 and a lower (higher) value of β_2 will produce a higher (lower) bias in the event-of-interest. These two biasing effects could cancel out to produce a situation similar to above. In our scenarios, we saw that, even when a treatment has no effect on the event-of-interest or a competing event (i.e. the treatment is a placebo), both a cause specific treatment effect and a subdistribution treatment effect can be found. This also implies that the biasing effect of unmeasured confounders (both omitted variable and correlation bias) can result in researchers reaching incorrect conclusions about how a treatment affects a population in multiple ways. We could have a treatment that is beneficial

for the prevention of both types of event, but due to the effects of an unmeasured confounder, it could be found to have a detrimental effect (for one or both) on patients from a subdistribution perspective.

Our investigation augments Lin et al's study into unmeasured confounding in a Cox model [145] by extending their conclusion (that bias is in the same direction as the confounder's effect and dependent on its strength) into a competing risks framework (i.e. by considering the Fine & Gray model as well) and demonstrating that this effect is reversed when there is confounding with the competing event. Lin et al. [145] also highlight the problems of omitted variable bias, which comes from further misspecification of the model; this finding was observed in our results as described above for Scenario 3.

The results from Scenario 7 (Weibull Distribution) and Scenario 8 (Plausible Distribution) are almost identical to those of Scenario 1 (No Effect) which implies that, by assuming both hazard functions in question are the same, we can assume they are both constant for simplicity. Since both the Cox and Fine & Gray models are ambiguous to underlying hazard functions and treatment effects are estimated without consideration for the baseline hazard function, it makes intuitive sense that the results would be identical regardless of what underlying functions were used to generate our data. This makes calculation of the explicit subdistribution treatment effect much simpler for future researchers.

Thompson et al. used the paradox that smoking reduces melanoma risk to motivate simulations similar to ours, which demonstrated how the exclusion of competing risks, when assessing confounding, can lead to unintuitive, mis-specified and possibly dangerous conclusions [162]. They hypothesised that the association found elsewhere [163] may be caused by bias due to ignoring competing events and used Monte Carlo simulations to provide examples of scenarios where these results would be possible. They demonstrated how a competing event could cause incorrect conclusions when that competing event is ignored - a conclusion we also confirm through the existence of bias induced on the Cox modelled treatment effect even with no correlation between the unmeasured confounder and treatment (i.e. $\theta_{\text{CSH},1} \neq 0$ in Scenarios 2 & 3). Thompson's team began with a situation where there may be a bias due to a competing event and reverse-engineered a scenario to find the potential sources of bias, whereas our study explored different scenarios and investigated the biased results they potentially produced.

Groenwold et al. [164] proposed methods to perform simulations to evaluate how much unmeasured confounding would be necessary for a true effect to be null given that an effect has been found in the data. Their methods can easily be applied to any metric in clinical studies (such as the different hazard ratios estimated here). Currently, epidemiologists will instigate methods such as DAGs, see Fig. 2.3, to visualise where unmeasured confounding may be a problem in analysis [165] and statisticians who deal with such models will use transition diagrams, see Fig. 2.1, to visualise potential patient pathways [113]. Using these two visualisation techniques in parallel will allow researchers to anticipate these issues, successfully plan to combat them (through changes to protocol or sensitivity analysis, etc. ...) and/or implement simulations to seek hidden sources of bias (using the methods of Groenwold [164] and Thompson [162]) or to adjust their findings by assuming biases similar to those demonstrated in our paper exist in their work.

The work presented here could be extended to include more complicated designs such as more competing events, more covariates and differing hazard functions. However, the intention of this paper was to provide a simple dissection of specific scenarios that allow for generalisation to clinical work. The main limitation of this work, to use of the same hazard functions for both events in each of our scenarios, was a pragmatic decision made to reduce computation time. The next largest limitation was the lack of censoring events, and was chosen to simplify interpretation of the model. This situation is unlikely to happen in the real world. However, since both the Cox and the Fine & Gray modelling techniques are robust to any underlying baseline hazard and independent censoring of patients [166], these simplifications should not have had a detrimental effect on the bias estimates given in this paper. This perspective on censoring is similar to the view of Lesko et al. [167] in that censoring would provide less clarity of the presented results.

2.5 Conclusion

This paper has demonstrated that unmeasured confounding in observational studies can have an effect on the accuracy of outcomes for both a Cox and a Fine & Gray model. We have added to the literature by incorporating the effect of confounding on a competing event as well as on the event-of-interest simultaneously. The effect of confounding is present and reversed compared to that of confounding on the event-of-interest. This makes intuitive sense as a negative effect on a competing event has a similar effect at the population level as a positive effect on the event-of-interest (and vice versa). This should not be overlooked, even when dealing with populations where the potential for competing events is much smaller than potential for the event-of-interest and is especially true when the two arms of a study are unequal. Therefore, we recommend that research with the potential to suffer from these issues be accompanied by sensitivity analyses investigating potential unmeasured confounding using established epidemiological techniques applied to any competing events as well as the event-of-interest. In short, unmeasured variables can cause problems with research, but by being knowledgeable about what we don't know, we can make inferences despite this missing data.

Blank Page

Chapter 3

Using Inverse-Probability-of-Censoring-Weights to Estimate Calibration-in-the-Large for Time-to-Event Models

MA Barrowman, A Pate, GP Martin, CJM Sammut-Powell, M Sperrin

Abstract

Introduction

A key component of the development of a prediction model/algorithm is the assessment of its calibration through means of validation (internal and external). For time-to-event models, this assessment is complicated in three ways:

- Calibration can be assessed at multiple time points,
- When Cox modelling has been used, there exists no "intercept" for a model to be assessed on
- Censoring occurs within the data, and this may or may not be correlated with the event-of-interest

We choose to focus on analysing methods of overcoming the third of these problems using Inverse Probability of Censoring Weighting (IPCW), which can also combat the other two problems.

Methods

We used simulations to generate time-to-event data with censoring, where censoring can be correlated or not with the event-of-interest. We then applied a pre-calibrated prediction models

(including flawed ones) to the data and assessed the calibrations of these models under different methods:

- Kaplan-Meier Method (KM), the KM curve is used as a comparator to the model predictions
- Logistic Regression with IPCW Weighting (LW) and without it (LU)
- Pseudo-Observations with IPCW (PW) and without it (PO), where the calibration is assessed using the pseudo-observations of the model data.

These simulations were aggregating and analysed to compare Bias and Coverage of each of these methods.

Results

The LU and PO Methods had increasing absolute Bias over time, regardless of whether the model was perfectly calibrated or not. At certain time points and in some scenarios, the PW method provided no bias, but rarely supplied an adequate level of coverage and was inconsistent over time. However, the LW Method consistently provided almost no Bias, and high coverage in all scenarios.

Discussion

The use of the IPCW, in combination with logisitic regression, when assessing calibration-in-the-large for time-to-event data provides the least biased method of assessment. This is in line with other work in this area that has shown similar results for calculation of the Brier Score and c-statistic.

Supplementary Material

Supplementary Material is available in Appendix C.

3.1 Introduction

Clinical prediction models (CPMs) are statistical models/algorithms that aim to predict the presence (diagnostic) or future occurrence (prognostic) of an event of interest, conditional on a set of predictor variables. Before they can be implemented in practice, CPMs must be robustly validated. Alongside assessment of overall performance (e.g. R^2 or the Brier Score) and discrimination (how well models discern between different patients), a fundamental test of a model's predictive performance is calibration; that is, the agreement between observed and predicted outcomes. This requires that among individuals with $p\%$ risk of an event, $p\%$ actually have the event across the full risk range [168]. The simplest assessment of calibration is the calibration-in-the-large, which tests for agreement in mean calibration (the weakest form of calibration) [169]. With continuous or binary outcomes, such a test is straight-forward: it can be translated to a test for a zero intercept in a regression model with an appropriately transformed

linear predictor as an offset, and no other predictors. More complicated measurements of calibration can also be assessed to describe how calibration changes across the risk range, such as calibration slope (see Appendix C).

In the case of time-to-event models, however, estimation of calibration is complicated in three ways. First, in the context of time-to-event models, one assesses the agreement between the observed and the estimated probability of the event occurring within a specified duration of time. Thus, calibration can be computed at multiple time-points, and one must decide which time-points to evaluate (and how to integrate over them [95]). The choice and combination of time-points determines what we mean by calibration; this is problem-specific and not the focus of this paper. Calibration can also be integrated over time using the martingale residuals [170]; however we focus on the case where calibration at a specific time point is of interest (e.g. as is common in clinical decision support).

Second, when a Cox proportional hazard model is used to fit the CPM, the baseline hazard function (and hence baseline survival) is not estimated per se [171]. The lack of baseline hazard can be overcome provided sufficient information concerning the baseline survival curve is available (although this is rarely the case as seen in QRISK [52], ASCVD [172] and ASSIGN [173]). Once this is established, estimated survival probabilities are available.

Third, censoring needs to be handled in an appropriate way and this is the focus of this paper. Censoring is commonly overcome by using Kaplan-Meier estimates [52], [171], but the censoring assumptions required for the Kaplan-Meier estimate are stronger than those required for the Cox proportional hazard model: the former requiring unconditional independence (random censoring), the latter requiring independence conditional on covariates only. This is a problem because when miscalibration is found using this approach, it is not clear whether this is genuine miscalibration or a consequence of the different censoring assumptions. Royston [174], [175] has proposed the comparison of KM curves within risk groups, which alleviates the strength of the independence assumption required for the censoring handling to be comparable between the Cox model and the KM curves (since the KM curves now only assume independent censoring within risk group). In these papers a fractional polynomial approach to estimating the baseline survival function (and thus being able to share it efficiently) is also provided. However, this does not allow calculations of the overall calibration of the model, which is of primary interest here.

QRISK used the overall KM approach in the 2007 paper [52] demonstrating adequate calibration (6.34% predicted vs 6.25% observed in women and 8.86% predicted vs 8.88% observed in men), but miscalibration in the QRISK3 update [5] (4.7% predicted v 5.8% observed in women and 6.4% predicted vs 7.5% observed in men). This may be because, as follow-up extends, the dependence of censoring on the covariates increases (QRISK had 12 years follow-up, QRISK3 had 18).

Royston [174] also presented an alternative approach for calibration at external validation, which utilises the approach of pseudo-observations, as described by Perme and Anderson [176] to overcome the censoring issue and produce observed probabilities at individual level. However, this assumes that censoring is independent of covariates. A solution to this problem is to apply a weighting to uncensored patients based on their probability of being censored according to a

model that accounts for covariates. The Inverse Probability of Censoring Weighting (IPCW) relaxes the assumption that patients who were censored are identical to those that remain at risk and replaces it with the assumption that they are exchangeable conditional on the measured covariates. The weighting inflates the patients who were similar to the censored population to account for those patients who are no longer available at a given time.

Gerds & Schumacher [177] have thoroughly investigated the requirements and advantages of applying an IPCW to a performance measure for modelling using the Brier score as an example and demonstrating the efficacy of its use, which was augmented by Spitoni et al [178] who demonstrated that any proper scoring rule can be improved by the use of the IPCW. This work has been extended by Han et al [179] and Liu et al [180] who demonstrated one can also apply IPCW to the c-statistic (a measure of discrimination).

In this paper we present an approach to assessing the calibration intercept (calibration-in-the-large) and calibration slope in time-to-event models based on estimating the censoring distribution, and reweighting observations by the inverse of the censoring probability. We compare simulation results from using this weighted estimate to an unweighted estimate within various commonly used methods of calibration assessment.

3.2 Methods

3.2.1 Aims

The aim of this simulation study is to investigate the bias induced by applying different methods of assessing model calibration to data that is susceptible to censoring and to compare it to the bias when this data has been adjusted by the Inverse Probability of Censoring Weighting (IPCW).

3.2.2 Data Generating Method

We simulated populations of patients with survival and censoring times, and took the observed event time as the minimum of these two values along with an event indicator of whether this was the survival or censoring time [154]. Each population was simulated with three parameters: β, γ and η, which defined the proportional hazards coefficients for the survival and censoring distributions and the baseline hazard function, respectively.

Patients were generated with a single covariate $Z \sim N(0, 1)$ from which, we then generated a survival time, T and a censoring time, C. Survival times were simulated with a baseline hazard $\lambda_0(t) = t^\eta$ (i.e. Weibull), and a proportional hazard of $e^{\beta Z}$. This allows the simulation of a constant baseline hazard ($\eta = 0$) as well as an increasing ($\eta = {}^1/_2$) and decreasing ($\eta = -{}^1/_2$) hazard function Censoring times were simulated with a constant baseline hazard, $\lambda_{C,0}(t) = 1$ and a proportional hazard of $e^{\gamma Z}$. Therefore, the hazard functions can be expressed in full as:

$$\lambda(t) = e^{\beta Z} t^\eta \qquad \lambda_C(t) = e^{\gamma Z}$$

This combines to give a simulated survival function, S as

$$S(t|Z=z) = \exp\left(-\frac{e^{\beta Z} t^{\eta+1}}{\eta+1}\right)$$

and a simulated censoring function, S_c as

$$S_c(t|Z=z) = \exp\left(-e^{\gamma Z} t\right)$$

Once the survival and censoring times were generated, the event time, $X = \min(T, C)$, and the event indicator, $\delta = I(T = X)$, were generated. In practice, only Z, X and δ would be observed.

During each simulation, we varied the parameters to take all the values,

- $\gamma = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$
- $\beta = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$
- $\eta = \{-\,{}^1/_2, 0,\ {}^1/_2\}$.

For each combination of parameters, we generated $N = 100$ populations of $n = 10,000$ patients (a high number of patients was chosen to improve precision of our estimates)

3.2.3 Prediction Models

For each population, we used three distinct prediction models for survival. F_P was chosen to exactly model the Data Generating Mechanism (DGM) to emulate a perfectly specified model:

$$F_P(t|Z=z) = 1 - \exp\left(-\frac{e^{\beta Z} t^{\eta+1}}{\eta+1}\right)$$

From this, we also derived a prediction model that would systematically over-estimate the prediction model, F_O, and one which would systematically under-estimate the prediction, F_U. These are defined as:

$$F_U(t|Z=z) = \text{logit}^{-1}\left(\text{logit}\left(F_P(t|z) - 0.2\right)\right)$$

$$F_O(t|Z=z) = \text{logit}^{-1}\left(\text{logit}\left(F_P(t|z) + 0.2\right)\right)$$

These prediction models were used to generate an estimate of the Expected probability that a given patient, with covariate z, will have an event at the given time. The value of 0.2 was chosen to provide a reasonably noticeable amount of under-/over-prediction.

3.2.4 The IPCW

In order to apply the IPCW, we need to calculate a censoring prediction model. For our purposes, we will again use a perfectly specified censoring distribution, G, to be derived directly from the DGM:

$$G(t|Z = z) = 1 - \exp\left(-e^{\gamma Z} t\right)$$

This is used to calculate an IPCW for all non-censored patients at the last time they were observed (t for patients who have not had an event, and X_i for patients who have had the event), This is defined as:

$$\omega(t|z) = \frac{1}{1 - G(\min(t, X_i)|z)}$$

Due to the sensitive nature of this propensity weighting over time, the IPCW is capped at a value of 10 [181], this capping should only effect around 2% of simulated patients.

3.2.5 Calibration Measurements

The prediction models were assessed at 100 time points, evenly distributed between the 25th and 75th percentile of observed event times, X. At each of these time points, we compare Observed outcomes (O) with the Expected outcomes (E) of the prediction models based on four choices of methodology [8], [52], [174], [175], [182] to produce measures for the calibration-in-the-large

- Kaplan-Meier (KM) - A Kaplan-Meier estimate of survival is estimated from the data and the value of the KM curve at the current time is taken to be the average Observed number of events within the population. The measure is the ratio of the Observed to the mean Expected number of events.

- Logistic Unweighted (LU) - Logistic regression is performed on the non-censored population to predict the binary Observed value using the logit(E) value as an offset and the Intercept of the regression is the estimate of calibration-in-the-large.

- Logistic Weighted (LW) - As above, but the logistic regression is performed using the IPCW as a weighting for each non-censored patient.

- Pseudo-Observations (PO) - The contribution of each patient (including censored patients) to the overall Observed value is calculated by removing them from the population and aggregating the difference. Regression is performed with the complimentary log-log function as a link function and the log cumulative hazard as an offset with the Intercept representing the estimate of calibration-in-the-large.

- Pseudo-Observations Weighted (PW) - As the PO method above, but the logistic regression is performed using the IPCW as a weighting in the same vein as the LW method.

The KM method is centred around 1 for well performing models, whereas the others are centered around 0, so we subtract 1 from the results of the KM method to centre it around the same value. Although this won't necessarily bring the KM estimate to the same scale, it will still allow for it to be compared to the others as it will still demonstrate when the KM measure is incorrect, show in which direction it is biased and give a magnitude relative to the confidence interval of how strong that bias is.

3.2.6 Estimands

For each set of parameters and methodology, our estimand at time, t, measured in simulation $i = 1, ..., N$ is $\theta_i(t)$, the set of estimates of the calibration-in-the-large for the F_P, F_U and F_O models in order. Therefore our underlying truth for all time points is

$$\theta = (0, 0.2, -0.2)$$

From this, we can also define our upper and lower bound for a 95% confidence interval as the vectors $\theta_{i,L}(t)$ and $\theta_{i,U}(t)$.

3.2.7 Performance Measures

The measures we will take as performance measures are the Bias, the Empirical Standard Error (EmpSE) and the Coverage (Cov). We will also estimate the MCMC Standard Error of each of these estimates in order to calculate a 95% confidence interval

Table 3.1: Performance Measures to be taken at each time point

Performance Measure	Estimation	SE
Bias	$\hat{\theta}(t) = \frac{1}{N}\sum_{i=1}^{N}\theta_i(t) - \theta$	$\hat{\theta}_{SE}(t) = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\theta_i(t) - \hat{\theta}(t)\right)^2}$
EmpSE	$\hat{E}(t) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\theta_i(t) - \hat{\theta}(t)\right)^2}$	$\hat{E}_{SE}(t) = \frac{\hat{E}(t)}{\sqrt{2(N-1)}}$
Coverage	$\hat{C}(t) = \frac{1}{N}\sum_{i=1}^{N} I\left(\theta_{i,L}(t) \leq \theta \leq \theta_{i,U}(t)\right)$	$\hat{C}_{SE}(t) = \frac{\hat{C}(t)\left(1-\hat{C}(t)\right)}{N}$

The bias provide a measures of how close our estimate is to the true value as per our data generating mechanisms. The coverage will demonstrate how often our confidence intervals surrounding our estimate actually include this true value. The Empirical Standard Error will show us how precise our estimates are.

3.2.8 Software

All analysis was done in `R 3.6.3` [183] using the various `tidyverse` packages [184], Kaplan-Meier estimates were found using the `survival` package [185]. Code to evaluate the pseudo-observations were taken from the `pseudo` R package [186], and adapted into C++ code using the `Rcpp` package [187] for faster implementation. The results app was developed using `shiny`[188]. The code used for this simulation study is available on Github and the results can be seen here

3.3 Results

Here, we present a subset of results with the full set of outputs available in the Calculator App. The estimates are presented with time on the x-axis and the y-axis showing the performance measure, stratified by model across facets and method of analysis by colour. We will investigate the Bias, EmpSE and Coverage for the scenarios where $\beta = 1$ and $\eta = {}^1/_2$ are fixed and γ varies

through -1, 0 and 1. These represent when the event and censoring are positively correlated ($\gamma = \beta = 1$), negatively correlated ($\gamma = -\beta = -1$) and when the covariate has no effect on the censoring distribution ($\gamma = 0$)

3.3.1 No correlation

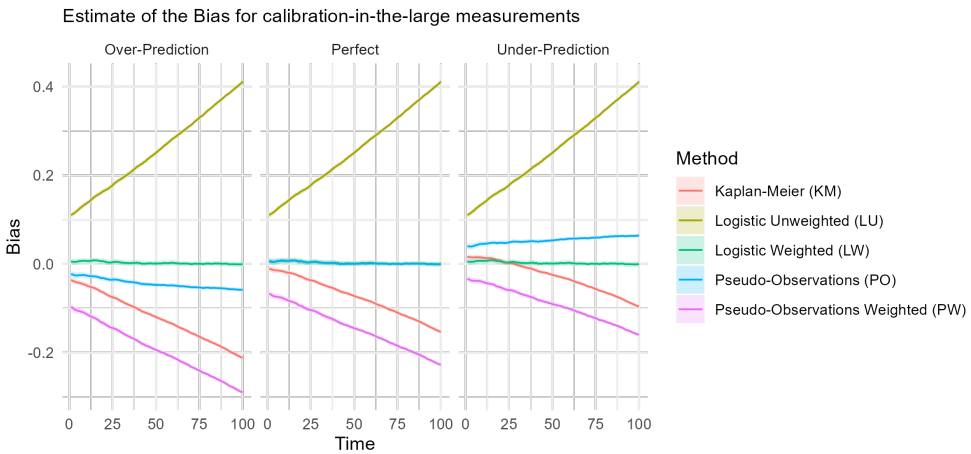

Figure 3.1: Bias for Over-estimating, Perfect and Under-estimating models across all four methods in the 'No correlation' scenario, when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

When $\gamma = 0$, we can see in figure 3.1 that the Bias is moving away from 0 for the LU and for the KM. For PO, it stays quite close to zero, but only touches for the Perfect model. For the PW method, the bias increases in magnitude over time for all models. The LW Method consistently provides an unbiased measurement of the model calibration regardless of the underlying accuracy of the model.

The small confidence intervals surrounding the Bias estimates above demonstrate that these results are consistent, which is exemplified by the small values found in figure 3.2. These results are also consistent across the three models.

The large biases above, also lead to inaccurate estimations at the simulation level with very low coverage of the true value, as demonstrated in figure 3.3. The LW Method once again is extremely close to the expected coverage of 95%, with PO matching similarly for the Perfect model.

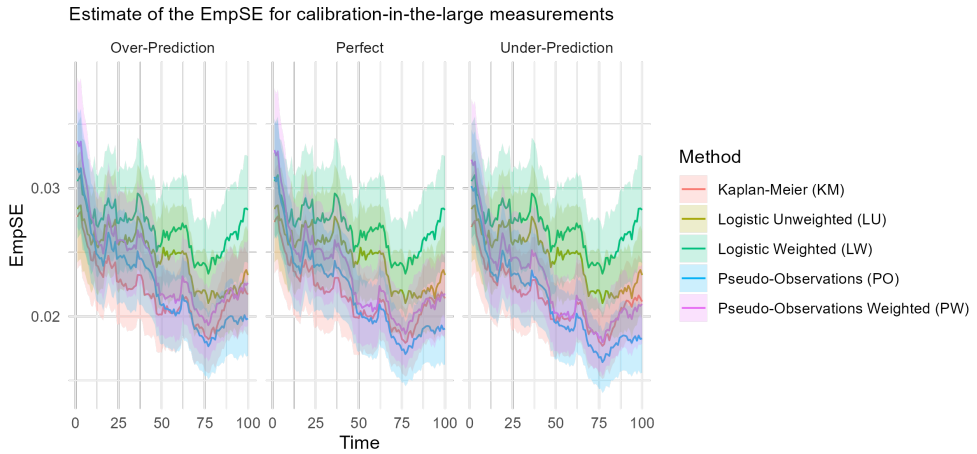

Figure 3.2: EmpSE for Over-estimating, Perfect and Under-estimating models across all four methods in the 'No correlation' scenario, when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

3.3.2 Positive correlation

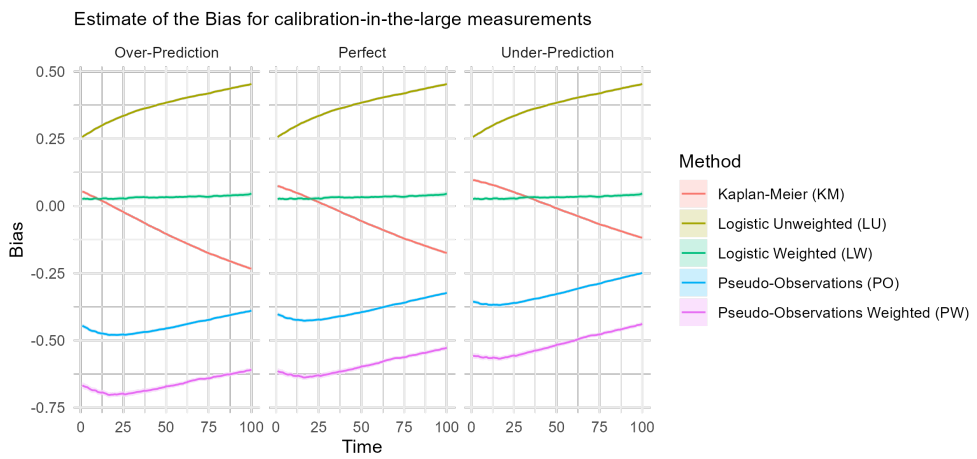

Figure 3.4: Bias for Over-estimating, Perfect and Under-estimating models across all four methods in the 'Positive correlation' scenario, when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

When $\gamma = 1$, we can in figure 3.4 see that the Biases for PO and LU, are once again very large with LU consistently over-estimating and PO and PW consistently under-estimating. KM touches 0 for the Perfect model, but not for the other two. LW is close to 0, but still has some small amount of positive bias. Some of the values shown here are quite extreme (as high as +0.5 bias and low as -0.7 bias).

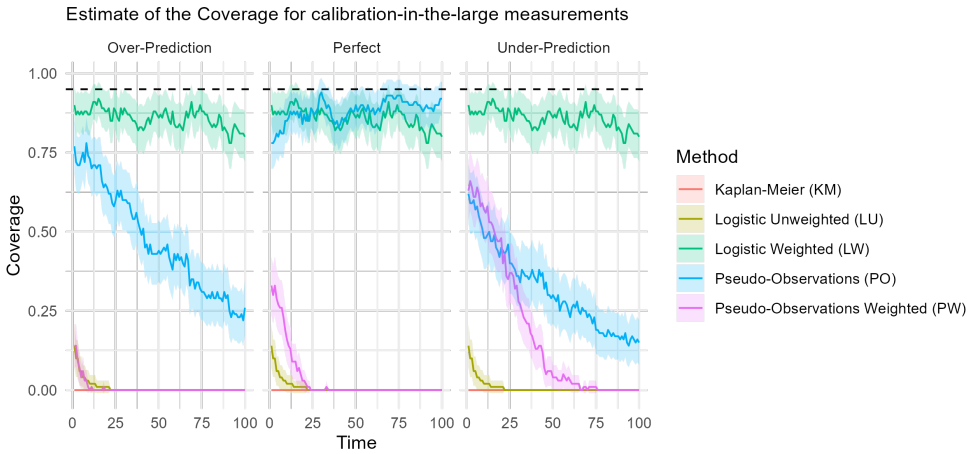

Figure 3.3: Coverage for Over-estimating, Perfect and Under-estimating models across all four methods in the 'No correlation' scenario, when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

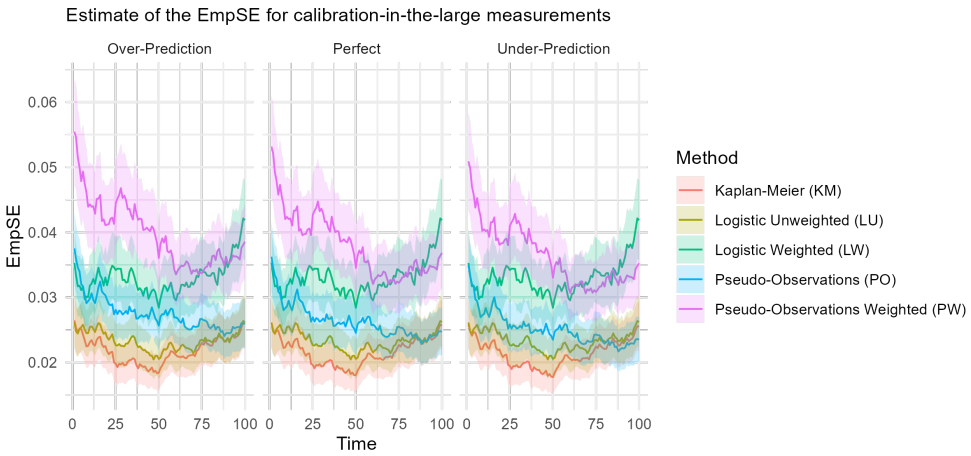

Figure 3.5: EmpSE for Over-estimating, Perfect and Under-estimating models across all four methods in the 'Positive correlation' scenario, when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

There is a larger EmpSE overall in this scenario than in the No correlation scenario, as seen in 3.5, with the PW method having the largest values for EmpSE showing a relatively high inconsistency across simulations, which may be expected with such a high magnitude of bias in these results.

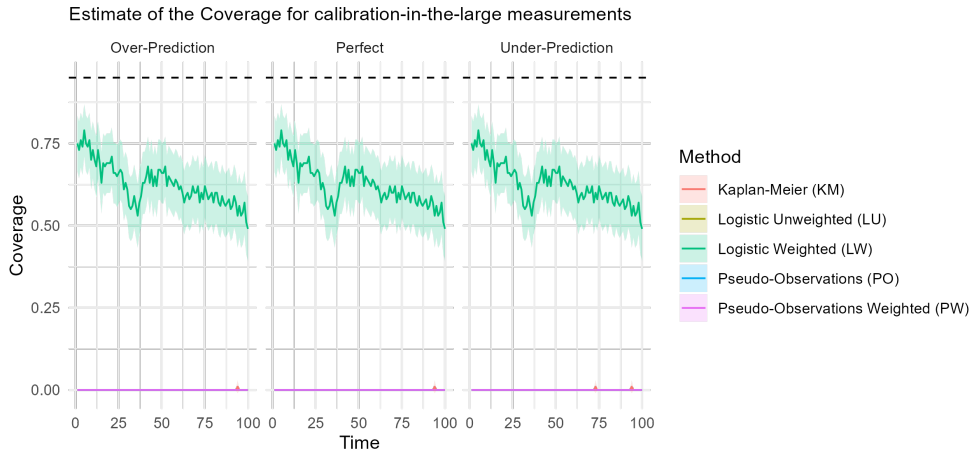

Figure 3.6: Coverage for Over-estimating, Perfect and Under-estimating models across all four methods in the 'Positive correlation' scenario, when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

However, there is an reduced level of coverage, even for the LW Method as shown shown in figure 3.6, with almost 0 coverage for non-LW methods.

3.3.3 Negative correlation

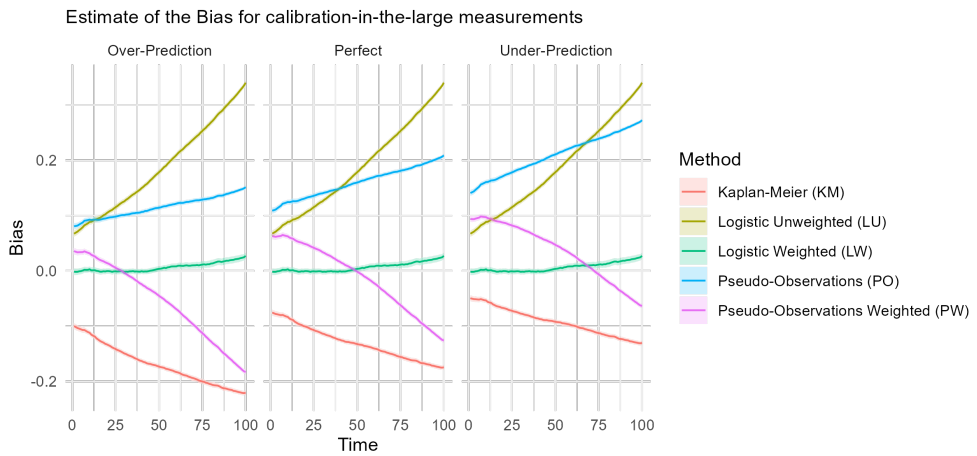

Figure 3.7: Bias for Over-estimating, Perfect and Under-estimating models across all four methods in the 'Negative correlation' scenario, when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

When $\gamma = -1$, we see can in figures 3.7 and 3.8, the results are similar in that LW remains close to 0 for all of the models, with the PW method crossing 0 at different points for the three models.

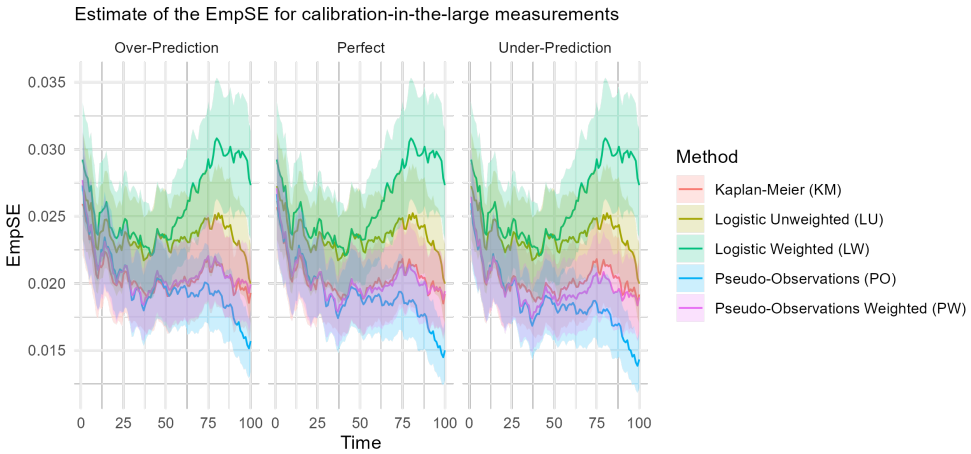

Figure 3.8: EmpSE for Over-estimating, Perfect and Under-estimating models across all four methods in the 'Negative correlation' scenario, when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

For the Negative correlation scenario, figure 3.8 shows us that the least consistent results can be found in the LW and PO methods as these are the highest across all three models, however the results still appear similar across the three models.

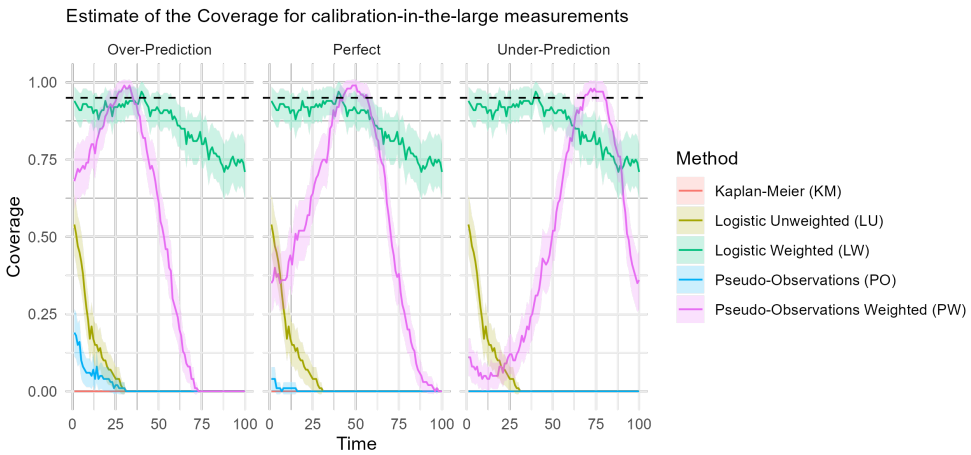

Figure 3.9: Coverage for Over-estimating, Perfect and Under-estimating models across all four methods in the 'Negative correlation' scenario, when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

The data shown in figure 3.9 demonstrates that the LW Method is again achieving high coverage, which dwindles only slightly over time. The PW spikes at the points where it crosses 0 in the bias plot above.

3.4 Discussion

This study has shown that using IPCW in combination with logistic regression can lead to unbiased estimates of calibration-in-the-large for survival models. Other methods that have been previously proposed, including Kaplan-Meier estimates and pseudo-observations, resulted in biased estimates of calibration-in-the-large, especially when the event and censoring distributions were correlated, even once the pseudo-observations method is adjusted by the IPCW.

Calibration is an important component of assessing the predictive performance of any prediction model. While methods of assessing calibration have been well-studied in CPMs for continuous or binary outcomes [8], [168], [189], methods in the context of time-to-event CPMs have received less attention. In practice, it is common to see time-to-event CPMs being developed using the Cox proportional hazards models, with calibration assessments made by comparing the predicted survival estimates with observed Kaplan-Meier estimates, such as in the validation of the QRISK model [52]. However, these two approaches make different assumptions regarding censoring, which leads to the bias in calibration-in-the-large that was observed in our simulations. Specifically, our simulations show that, even for a perfectly specified model, the Kaplan-Meier estimate of calibration-in-the-large is biased, especially as follow-up time increases. Hence, this method should be avoided since it makes it impossible to differentiate genuine miscalibration, from artificial miscalibration induced through violations of assumptions surrounding censoring.

In contrast, we found that estimates of calibration-in-the-large at a given time for time-to-event CPMs should be based on the logistic regression IPCW method in the presence of right-censored survival data in the validation cohort. The use of IPCWs removed bias and provided suitable coverage in the majority of simulation scenarios we considered. It should be noted that this high level of coverage, may be due to there being wider confidence intervals in the LO & LW methods as these are only applied to the uncensored patients remaining in the population at each given time point, which may also lead to a loss of information if censoring is particularly high. However, the low-levels of bias and EmpSE demonstrate that, even with the wider confidence intervals the results are still viable.

This supports previous research in this space, which has shown the validity of applying an IPCW to estimate Brier score and c-statistics in time-to-event models [177]–[180]. To our knowledge, no previous study has looked at the use of IPCW in the context of calibration assessment. Advantageously, using IPCW only assumes that censoring and event distributions are exchangeable conditional on the measured covariates.

We note that previous studies have proposed graphical methods of assessing calibration through smoothed calibration curves [95], which could be considered alongside the methods described here. Such smoothed calibration curves are produced by regressing the log-hazard of the outcome as a (smooth) function of the complementary log-log transformation of the predicted outcome probability at a set time, which can be summarised using an integrated calibration index [95]. Combining numeric summaries of calibration-in-the-large and calibration slope, with graphical calibration assessments would create a strong form of calibration assessment for time-to-event CPMs [169].

Importantly, another key issue with assessing calibration of time-to-event CPMs is that one

needs absolute estimates of survival probabilities. Given that many time-to-event CPMs are based on the Cox proportional hazards model, this is not (usually) reported. While this was not the focus of the current paper, metrics of strong calibration, upon independent validation, do require the baseline survival function to be reported [171]. This is also the case when using the IPCW method. For this reason, developing time-to-event CPMs based on parametric or flexible parametric models is usually more appropriate than Cox proportional hazard models [67].

Some limitations should be noted when interpreting the results of this study. Firstly, we only investigated the methods within a simulation study. This is required so we could evaluate bias of each method in estimating calibration-in-the-large.

Secondly, throughout the simulations, the model used to calculate the IPCWs was perfectly specified. In practice, one would not know the "true" censoring distribution, and so care would need to be taken in specifying the IPCW model in order to derive accurate weights. This would involve using well-known and standardised model derivation techniques (such as model selection, etc...) to ensure that the censoring model used to calculate the IPCWs is as accurate as possible.

Further studies could be use to assess the use of the graphical methods mentioned above in comparison to those studied here, or to investigate the effect that higher censoring rates might have on the results. This is prescient since LW & LO methods are applied only to the non-censored patients and thus a high censoring rate could lead to a large change in these results, and have the effect of massively widening the confidence intervals.

In conclusion, this paper has shown, through a comprehensive simulation study, that only the IPCW method produced unbiased estimates of calibration-in-the-large for time-to-event CPMs when there is censoring within the validation set. The commonly applied use of Kaplan-Meier estimates was highly biased in most scenarios. These results suggest that the calibration-in-the-large at a given time of a time-to-event CPM should be evaluated by fitting a weighted logistic regression model to the observed binary event indicator at that time, with the linear predictor of the model as an offset; the weights should be estimated using the IPCW method.

Chapter 4

Development of Predictive Performance Metrics for the Validation of Multi-State Clinical Prediction Models

MA Barrowman, GP Martin, N Peek, M Lambie, M Sperrin

Abstract

Introduction

Multi-State Clinical Prediction Models (MSCPMs) can provide a multi-dimensional estimate of patint outcomes over time. However, there currently does not exist any viable or robust manner in which to validate these models.

Methods

We combine current methods of assessing multinomial outcomes and the Inverse Probability of Censoring Weighting, which adjusts patient populations for time-dependent outcomes where censoring is present. We provide methods of investigating the overall Accuracy, Calibration and Discrimination of an MSCPM and provide insights into their interpretation. These extensions are:

- Accuracy - The Multiple Outcome Brier Score
- Discrimination - The Polytomous Discriminatory Index
- Calibration - Multinomial Calibration Intercept, Matched Slope and Unmatched Slope.

Measurements can be taken at any time through the use of the ICPW, and we provide methods of standardising these measures to be on the same scale as the traditional metrics that they are extensions of.

Results

We applied these methods to an MSCPM developed in Chronic Kidney Disease, and compare them to the traditional metrics, which are designed for a two-state system (as in traditional survival analysis). Results show that the new metrics line up with the traditional ones and provide an overall summary.

Discussion

The results of applying our novel methods line up well with the traditional methods and demonstrate that they are viable summary statistics for assessing the performance of an MSCPM. They can be compared directly regardless of the shape of or number of states in the model, including to traditionally developed time-to-event models.

Supplementary Material

Supplementary Material is available in Appendix D.

4.1 Introduction

Clinical Prediction Models (CPMs) provide individualised risk of a patient's outcome [8], based on that patient's set of predictor variables. These predictions will often be in the form of a risk score or probability, but can also be expressed as expected values if the outcome is continuous. However, using traditional modelling techniques, these CPMs will only predict a single outcome. There are an increasing number of clinical applications where predicting the risks of a multi-dimensional outcome is of interest, this can be through a multinomial prediction where time is fixed or as a sequence of outcomes over time. For example, in Chronic Kidney Disease, there is an interest in being able to predict the probability of receiving RRT prior to death. To this end, Multi-state Clinical Prediction models (MS-CPMs) provide the framework in which to do this.

Once a CPM has been developed, it is important to assess how well the model actually performs [168]. This process is called Model Validation and involves comparing the predictions produced by the model to the actual outcomes experienced by patients. It is expected that the development of a CPM will be accompanied by the validation of the model on the same dataset it was developed in (internal validation), using either bootstrapping or cross-validation to account for optimism in the developed model [17]. Models can also be validated on a novel dataset (external validation), which is used to assess the generalisability and transportability of the model [190].

During validation, there are different aspects of model performance that we can assess and these are measured using specific metrics, for example, to assess the overall Accuracy

of a model, we may use the Brier Score [191] . In the field of clinical prediction modelling, the *Discrimination* of a model gives a measure of how well the model differentiates between patients (e.g. those who have had the outcome versus those who have not) and the *Calibration* gives a measure of how close the individual predicted values are to the outcomes. The current metrics that are commonly used have been designed and extended to work in a variety of model development frameworks. However, these measures have not been applied to a multi-state context and thus would need to account for the sequence of outcomes over time as well as for the censoring of patients (as common occurs in survival data). This paper aims to provide use-able extensions to current performance metrics to be used when validating MS-CPMs. It is essential that these extensions are directly comparable with current metrics (to allow for quicker adoption), that they reduce to the current metric in relevant simple cases and that they appropriately account for the censoring of patients.

Currently, the most common way to validate an MS-CPMs is by applying traditional methods to compare across two states at a given time and then aggregating the results in an arbitrary manner. Other methodologists have extended existing metrics to multinomial outcomes [192], which do not contain a time-based component; to simple competing risks scenarios [193], which do not contain transient states; or to time dependent outcomes [194]; which do not have multiple states. Spitoni et al [178] developed methods to apply the Brier Score (or any proper score functions) to a multi-state setting and so a simplified and specific version of their work is described in this paper. We also combine these ideas to apply an extension of the c-statistic (which measures discrimination) and an extension of the logistic intercept and slope (which measure calibration) to the multi-state context.

The aim of this chapter is to develop metrics to evaluate the predictive performance of a Multi-State CPM, building upon traditional metrics (and extensions thereof) where appropriate. In Section 4.2, we will introduce a motivating example and dataset. In Section 4.3 we will define the traditional metrics used for assessing accuracy, calibration and discrimination and how they can be applied to a time-to-event model and how we can account for censoring in these metrics. In Section 4.4 we will build on these metrics and extend them into a Multi-State framework and then apply all these methods to our motivating example in Section 4.5

4.2 Motivating Data Set

Throughout this paper we will use a model developed in Chronic Kidney Disease (CKD) patients to assess their progression onto Renal Replacement Therapy (RRT) and/or Death. For illustration, we develop an MS-CPM for this context using data from the Salford Kidney Study (SKS) and then externally validated it using data from the West of Scotland (see Table 4.1). This mimics the process taken to develop such a model for clinical use (as in Chapter 5).The original model predicts the probability that a patient has begun RRT and/or died after their first recorded eGFR below 60 ml/min/1.73m^2, by any time in the future (reliable up to 10 years). For the purposes of this paper, we will take a "snapshot" of the predictions at the 5 year time point, however these methods can work at any time point, and by assessing at multiple time-points, users are able to create a smooth estimate for these values and assess how

Table 4.1: Population demographics, continuous displayed as median (Inter-Quartile Range), and Categorical/Comorbidity data as number (percent) range and number missing are also included

Variable	median (IQR) or n (%)	range	missing (%)
Age	69.000 (17.000)	[11.000, 98.000]	0 (0.00%)
eGFR	35.398 (20.565)	[1.199, 59.994]	0 (0.00%)
eGFR Rate	-0.932 (21.897)	[-28.636, 50.653]	0 (0.00%)
SBP	144.000 (30.000)	[82.000, 258.000]	6345 (82.31%)
DBP	77.000 (17.000)	[35.000, 128.000]	6345 (82.31%)
BMI	28.081 (6.811)	[17.073, 52.403]	7491 (97.17%)
Albumin	37.000 (6.000)	[7.000, 53.000]	3134 (40.65%)
Calcium	2.410 (0.160)	[1.455, 3.400]	4513 (58.54%)
Haemoglobin	116.000 (29.000)	[7.100, 208.000]	3557 (46.14%)
Phosphate	1.160 (0.320)	[0.320, 4.370]	4510 (58.50%)
uPCR	0.062 (0.177)	[0.001, 1.943]	7170 (93.01%)
uPCR Rate	0.004 (0.905)	[-176.200, 952.812]	7495 (97.22%)
Gender:			0 (0.00%)
Male	3885 (50.40%)		
Female	3824 (49.60%)		
Ethnicity:			7009 (90.92%)
White	679 (8.81%)		
Asian	12 (0.16%)		
Black	7 (0.09%)		
Other	2 (0.03%)		
Smoking Status:			7709 (100.00%)
Former	0 (0.00%)		
Non-Smoker	0 (0.00%)		
Smoker	0 (0.00%)		
Former (More than 3Y)	0 (0.00%)		
Diagnosis Group:			6640 (86.13%)
Systemic diseases affecting the kidney	316 (4.10%)		
Glomerular disease	278 (3.61%)		
Tubulointerstitial disease	173 (2.24%)		
Miscellaneous renal disorders	198 (2.57%)		
Familial / hereditary nephropathies	104 (1.35%)		
Comorbidities:			
CCF	408 (5.29%)		0 (0.00%)
COPD	0 (0.00%)		7709 (100.00%)
CVA	186 (2.41%)		0 (0.00%)
DM	1535 (19.91%)		0 (0.00%)
HT	3114 (40.39%)		0 (0.00%)
IHD	863 (11.19%)		0 (0.00%)
LD	0 (0.00%)		7709 (100.00%)
MI	556 (7.21%)		0 (0.00%)
PVA	376 (4.88%)		0 (0.00%)
ST	0 (0.00%)		7709 (100.00%)

the measures may change over time (i.e drift). The predictions and metrics applied to them will depend on the population distribution at the 5-year snapshot, which can be seen in Table 4.2. For brevity, this table uses acronyms for the comorbidities which can be seen below:

- CCF - Congestive Cardiac Failure
- COPD - Chronic Obstructive Pulmonary Disease
- CVA - Prior Cerebrovascular Accident
- DM - Diabetes
- HT - Hypertension
- IHD - Ischemic Heart Disease

Table 4.2: Distribution of patients in Population at 5 years including number of times each transition occured

State	Pop	Prop	->RRT	->Death	->Cens
CKD	4,391	67.81%	742	1,427	1,144
RRT	458	7.07%		199	85
Death	1,626	25.11%			

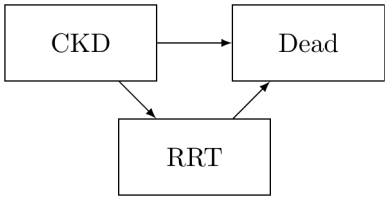

Figure 4.1: Layout of the MSM used in the motivating model

- LD - Chronic Liver Disease
- MI - Prior Myocardial Infarction
- PVA - Peripheral Vascular Disease
- ST - Prior Solid Tumour

The Three-State model used in our example is designed as an Illness-Death Model [113], this is one of the simplest MSM designs and has the key advantage over a traditional model that they can predict whether a patient is in or has visited the transient state before reaching the absorbing state (i.e. patient who became ill before dying or who started RRT before dying), see figure 4.1. It is important to note that the shape of this model is used for illustrative purposes only and the methods described in this paper can be used for *any* shape of MS-CPM For the purposes of this paper, we will not discern between whether a patient in the Death state has visited RRT or not, and so our model has $K = 3$ states. If we were to differentiate between these two, then the "Death" state would essentially be split in two, and so the model would have four states to describe the four pathways a patient can take (and so $K = 4$).

We will calculate the Multi-State Model Extension metrics for the model developed here, as well as the defined confidence intervals around these estimates.

4.3 Current Approaches and Preliminaries

In this section, we describe three commonly used performance metrics for assessing the performance of a traditional survival clinical prediction model and the conventions and notation used throughout this paper. These metrics assess the Accuracy, Discrimination and Calibration of the models being validated. Accuracy is an overall measurement of how well the model predicts the outcomes in the patients. Discrimination assesses how well the model discerns between patients; in a two-state model this is a comparison of patients with and without the outcome, and should assign a higher value to those that experience the outcome. Calibration is the agreement between the observed outcomes and the predicted risks across the full risk-range.

We are applying cross-sectional metrics at a set time point within the setting of a time-to-event model and so we need to account for the censoring of patients and therefore, each uncensored patient at a given time t will be weighted as per the Inverse Probability of Censoring Weighting (IPCW) [195]. This allows the uncensored patient population to be representative of the entire patient population.

4.3.1 Baseline Models

To assess the performance of a model, we must compare the values produced by the performance metrics to those of two baseline models; a random or non-informative model and a perfect model.

A Non-Informative (NI-)model assigns the same probability to all patients to be in any state regardless of covariates and is akin to using the average prevalence in the entire population to define your model. For example, in a Two-State model with an event that occurs in 10% of patients, all patients are predicted to have a 10% chance of having the event. For many metrics, models can be compared to an NI-model to assess whether the model is in fact "better than random."

A Perfect (P-)model is one which successfully assigns a 100% probability to all patients, and the predictions are correct; this is the ideal case and is therefore the standard that most models aim for.

The metrics produced by these baseline models will often depend on the prevalence of each state and/or the number of states. These values can be used as comparators to provide contextual information regarding the strength of model performance. These baselines metrics for the NI-model and the P-model will be referred to as the NI-level and P-level for the metric.

4.3.2 Notation

Throughout this paper, we will use consistent notation which is shown here for reference and to avoid repetition in definitions. The common notations are defined below: Other notation will be define as they are introduced.

4.3.3 Patient Weighting

At a given time after the index date, some patients in our validation data set will be censored and so our performance metrics must adjust for this. Therefore, all patients will be subject to IPCW, which applies a higher weighting to patients who are more likely to be censored. This process is assumed to be conditionally independent of the Multi-State process, given a patient's covariates [178].

To calculate this weight, first we need to estimate an individual patient's probability of not being censored at the current time point, $G(t|Z)$, where Z is the patient's covariate characteristics and t is the current time point. This is done in our validation cohort using a Cox regression which provides estimated hazard ratios for each of the covariates $\hat{\beta}$ taking the time of censoring as the event-of-interest. Absolute predictions are then calculated using the Breslow estimate of

Table 4.3: Common Notation used throughout this paper

Notation	Meaning
$N(t)$ or N	Number of (non-censored) patients in a population at time t
K	Number of states predicted by the model
$P_i^k(t)$ or P_i^k	Predicted probability of whether patient i was in state k at time t
$P_i^{!k}(t)$ or $P_i^{!k}$	Predicted probability of whether patient i was not in state k at time t, i.e. $P_i^k + P_i^{!k} = 1$
$P_i(t)$ or P_i	If $K \neq 2$, vector of predicted probabilities for patient i at time t, $P_i = (P_i^1, P_i^2, ..., P_i^K)$ If $K = 2$, then $P_i = P_i^2$ (i.e. predicted probability of the second state at time t)
$P^k(t)$ or P^k	The vector of the predicted probabilities of being in state k for the whole population at time t
$P(t)$ or P	If $K \neq 2$, a $N \times K$ matrix of predicted probabilities for each state & individual at time t If $K = 2$, a vector of the predicted probabilities of being in state 2 for the whole population at time t
$O_i^k(t)$ or O_i^k	Binary indicator for whether patient i was in state k at time t
$O_i^{!k(t)}$ or $O_i^{!k}$	Binary indicator for whether patient i was not in state k at time t, i.e $O_i^k + O_i^{!k=1}$
$O_i(t)$ or O_i	If $K \neq 2$, vector of outcomes for patient i at time t, $O_i = (O_i^1, O_i^2, ..., O_i^K)$ If $K = 2$, then $O_i = O_i^2$ (i.e. observation of patient in the second state at time t
$O^k(t)$ or O^k	The vector of observed outcomes of being in state k for the whole population at time t
$O(t)$ or O	If $K \neq 2$, a $N \times K$ matrix of observed proportions for each state & individual at time t If $K = 2$, a vector of the observed proportions for state 2 for the whole population at time t
$Q^k(t)$ or Q^k	The weighted proportion of the population in state k at time t
$Q(t)$ or Q	The vector of weighted proportions of the population in all states at time t, $Q = (Q^1, Q^2, ..., Q^K)$
$\omega_i(t)$ or ω_i	Weighting given to patient i at time t
$\omega(t)$ or ω	The vector of weights given to entire population at time t
$N_\omega(t)$ or N_ω	Weighted size of population at time t, $N_\omega = \sum_{i=1}^{N} \omega_i$
$N_k(t)$ or N_k	Weighted size of state k at time t, $N_k = \sum_{i \in A_k} \omega_i$

the cumulative baseline hazard function, $\hat{\Lambda}_0$. The estimate, $\hat{G}$, is then given by

$$\hat{G}(t|Z) = \exp\left(-e^{\beta Z}\hat{\Lambda}_0(t)\right)$$

For a given patient, i, with a maximum observed time of T_i, we will define $\delta_i = 0$ if the patient was censored and $\delta_i = 1$ if the patient moved to an absorbing state (e.g. died) and z_i to be that patient's set of covariates.

We can therefore define the IPCW for patient i at time t to be:

$$\omega_i(t) = \frac{I(T_i \leq t_i, \delta_i = 1)}{\hat{G}(T_i|z_i)} + \frac{I(T_i > t_i)}{\hat{G}(t_i|Z_i)}$$

By applying this weight to all patients included at each time point under analysis, we can be confident that our measurements are robust to right-censored data, subject to the assumptions made in their definition.

The metrics defined below (including those traditionally defined elsewhere) have been corrected for the effect of censoring by applying the IPCW, $\omega_i(t)$ to each patient as a multiplicative weight.

4.3.4 Accuracy - Brier Score

For these metrics, we will be considering predictive performance of the models at specific time point($t = 5$ years in our CKD example), and so we simplify notation by removing the references to time given above, for example $\omega_i = \omega_i(5 \text{ years})$.

The Brier Score is used to assess the overall accuracy of predictions, which assigns a score to each observation dependent on the predicted probability and the outcome. It then averages these scores across the entire population. The Brier Score, adjusted for IPCW, for a single outcome model for a single patient is given by:

$$\text{BS}_i = \omega_i \left(P_i - O_i\right)^2$$

And for the entire population, we take the weighted average given by the following [191]

$$\text{BS} = \frac{1}{N_\omega}\sum_{i=1}^{N}\text{BS}_i = \frac{1}{N_\omega}\sum_{i=1}^{N}\omega_i \left(P_i - O_i\right)^2$$

A lower Brier score implies a more accurate model (since the Predictions and the Observations will be closer to one another). The P-level of the BS measure is 0 and the NI-level is $Q(1-Q)$.

In order to standardise the Brier Score, we can rescale it by dividing by the NI-level and subtracting it from 1 to give the adjusted Brier Score (aBS):

$$\text{aBS} = 1 - \frac{BS}{Q(1-Q)}$$

The aBS brings the NI-level to 0 and the P-level to 1 and so a higher value for the aBS implies a model accurate model. One thing to note is that it is possible to get negative values for the

aBS if a model performs worse than a non-informative model; however in practice this model would essentially be unusable as it is (although still useful if predictions were reversed).

We can use the values of BS_i to calculate a standard deviation and thus build a confidence interval surrounding our overall BS estimate by use of the relevant z-score and assuming the underlying distribution of possible BS scores follow a Normal distribution [196], [197]. This population-based BS confidence interval can be converted into a confidence interval for the aBS using the above formula.

4.3.5 Discrimination - c-statistic

The c-statistic [198] is the most common method to assess the discriminative ability of a prediction model. In a traditional model, at a single time point (cross-sectional or not), this can be interpreted as the probability that two patients, chosen at random from the two outcome groups, will be correctly discriminated. Here, correct discrimination means that the patient who had the event was predicted to have a high probability of having the event than the patient who did not have the event.

$$c = \mathrm{Prob}\left(P_i < P_j \mid O_i = 0 \;\&\; O_j = 1\right)$$

This can be estimated empirically by averaging over all pairs of patients where one is selected from each state:

$$\hat{c} = \frac{1}{N_1 N_2} \sum_{i \in A_1} \sum_{j \in A_2} \omega_i \omega_j C_2(P_i, P_j)$$

where

$$C_2(a,b) = \begin{cases} 1 & a < b \\ 0 & a > b \\ \frac{1}{2} & a = b \end{cases}$$

In practice, it will be very rare for two predicted probabilities to be exactly equal, but this case is needed to account for the NI-model and produce the NI-level of 0.5, we also have a P-level of 1 regardless of the prevalence of the two states.

We can calculate the variance, therefore the standard error confidence intervals, around this estimate for $\hat{c}$ by using the method as described by DeLong et al [199], [200]:

$$\mathrm{var}\left(\hat{c}\right) = \frac{1}{N_1} S_1 + \frac{1}{N_2} S_2$$

where S_1 and S_2 are given by the following:

$$S_1 = \frac{1}{N_1 - 1} \sum_{i \in A_1} \left(V_1(i) - \hat{c}\right)^2 \qquad S_2 = \frac{1}{N_2 - 1} \sum_{j \in A_2} \left(V_2(j) - \hat{c}\right)^2$$

and V_1 and V_2 are:

$$V_1(i) = \frac{1}{N_2} \sum_{j \in A_2} \omega_i \omega_j C_2(P_i, P_j) \qquad V_2(j) = \frac{1}{N_1} \sum_{i \in A_1} \omega_i \omega_j C_2(P_i, P_j)$$

4.3.6 Calibration - Intercept and Slope

In a traditional model, the Calibration Intercept is a measure of Calibration-in-the-Large, or overall calibration across the entire population [11]. Calibration slope indicates how well the model predicts across different prediction values. These metrics can be measured using logistic regression on the probability of the outcome using the logit of the prediction as the predictor in the regression:

$$\mathrm{E}\left[\mathrm{logit}\left(O\right)\right] = \alpha + \beta \mathrm{logit}\left(P\right)$$

The estimates of these coefficients, $\hat{\alpha}$ and $\hat{\beta}$ are found using a weighted binomial logistic regression, with weights ω_i. The intercept, $\hat{\alpha}$, can provide a measure of any systemic over- or under-prediction of the outcome within the model. The slope, $\hat{\beta}$, provides a measure of how well the model performs across the population, rather than simply an average of the population (as $\hat{\alpha}$ is). It is advised that the intercept is calculated on its own first using logit(P) as an offset (without a predictor, i.e. fixing $\beta = 1$) and then the slope is calculated using $\hat{\alpha}$ as an offset [8]; however, for simplicity and consistency with the MSM metric we have chosen to model them both together.

As the predicted values of an NI-Model would be the same for all patients, a directly calculated NI-model would not converge, however the limit of such a model (as the individual predictions tend to equality) would give NI-levels for the Intercept equal to prevalence (Q) and slope equal to 0 (since every subgroup has the same predicted value). For a P-model, the Intercept would be 0 and the slope would be 1.

These metrics, intercept and slope, are usually described with an interpretation depending on the fit and whether the P-level (0 and 1, respectively) is within the confidence interval and, if not, which direction the miscalibration lies. If the calibration intercept is considered to be above or below the P-Level, then it indicates that the model is systemically under- or over-predicting the results, respectively. Similarly, a calibration slope that is below or above the P-Level is interpreted to mean that the model had predictions that were too extreme or too moderate across the prediction spectrum [190].

4.4 Extension to Multi-State Models

4.4.1 Trivial Extensions

As well as the extension methods described in this paper, each of the traditional performance metrics described above can be applied to a MS-CPM with trivial extension. These require the predictions and outcomes to be reduced to a model with only two states which allows the traditional performance metrics to be directly applied.

The first method, One-Vs-All, is based on whether a patient is in each state or not at a

given time. For each state, we take the current state as the outcome state and collapse all other states into a single "not-" state. For example, when analysing the CKD state, we collapse RRT and Death into a single "not-CKD" state. This gives us a metric for each state in the model.

The second method, Pairwise, compares across pairs of states by ignoring predictions unrelated to them at a given time. For each pair of states, we exclude patients not in one of the two states and normalise the two predicted probabilities so that they sum to 1. For example, when assessing CKD vs RRT, we exclude all patients in the Death state, take our outcome state as RRT and divide the predicted probability of being in RRT by the predicted probability of being in either CKD or RRT (i.e. probability of being in RRT given that they are in either RRT or CKD). This gives us a metric for each pair of states in the model.

The third method, Transition-wise, compares patients undergoing a specific transition. We take the subset of patients who were eligible for a transition and classify those who underwent the transition as being in the outcome state and compare them to those that didn't undergo the transition (by the given time). In our example, when looking at the RRT to Death transition, we would take the subset of all patients who underwent the CKD to RRT transition (i.e. those eligible for the RRT to Death transition) and compare those who transitioned to Death with those who remained in the RRT state.

Note that the subset of patients in the second and third methods are not always equivalent. When analysing RRT to Death or RRT vs Death, the patients in the RRT state are the same, but the patients in the Death state are different (RRT vs Death includes those that went directly from CKD to Death). The predicted probabilities are similarly different.

However, the above collapsing methods are a simplistic representation of the MS-CPM, and it would be more informative to consider the performance of the different state-transitions. To achieve this, we here propose novel extensions to each performance metric, using the methods outlined in section 1.3 as the foundation. Throughout, we consider the evaluation of predictive performance at a single follow-up time (e.g. t=5 years in our CKD example). As such, we base our methodological development on the methods designed for multinomial outcomes [201] , and here we are going to use a similar idea for validating a multi-state model. The key novelty is that we combine ideas from validation measures of time to event, such as IPCW to those for multinomial models/outcomes .

4.4.2 Accuracy - Multiple Outcome Brier Score

Brier's original definition of the Brier Score [191] was designed to assess predictions of multinomial outcomes, particularly in weather forecasting. By adapting this model and applying the time-dependent IPCW to each individual, their Multiple Outcome Brier Score can be calculated as:

$$\mathrm{BS}_{i,K} = \omega_i \left(\sum_{k=1}^{K} \left(P_i^k - O_i^k \right)^2 \right)$$

We then take an average to find the overall BS:

$$\mathrm{BS}_K = \frac{1}{N_\omega}\sum_{i=1}^{N}\mathrm{BS}_{i,K} = \frac{1}{N_\omega}\sum_{i=1}^{N}\sum_{k=1}^{K}\omega_i\left(P_i^k - O_i^k\right)^2$$

This version of the Brier Score is a generalisation of the traditional Brier Score, which can be applied to multiple outcomes and accounts for patient censoring. Similar to the traditional Brier Score, a lower value implies a more accurate model. If the two Brier Score measures are applied to a Two-State Model, then the generalised BS above is twice that of the traditional BS, ($\mathrm{BS}_K = 2\mathrm{BS}$), this is because the traditional metric looks at only the outcome state, but the extended method sums over both states, which will score identically.

For this metric, the P-level is 0 and, similar to the traditional metric, the NI-Level is $\sum_{k=1}^{K} Q_k(1-Q_k)$; because of this, we would need to apply an adjustment similar to the traditional Brier Score:

$$\mathrm{aBS} = 1 - \frac{\mathrm{BS}_K}{\sum_{k=1}^{K} Q_k(1-Q_k)}$$

Note that due to the relationship between BS and BS_2, the doubling that occurs cancels out between the numerator and denominator and so this adjustment works on the same scale as the previously defined aBS (and thus is given the same name).

As with the traditional BS metric, each patient will have their own BS_K measurement and so we can find the population-based confidence interval for the BS_K by using the standard deviation of these values. This can once again be converted into a confidence interval for the aBS.

4.4.3 Discrimination - Polytomous Discriminatory Index

Intuitively, the extension of the c-statistic would be the probability that K patients, chosen randomly from each of the outcome groups, will be correctly discriminated. In this case, what it is to be correctly discriminated needs to be defined. The Polytomous Discriminatory Index (PDI) provides a definition for this discrimination [193]. We define a K-tuple of patients as an ordered set of K patients where one patients is from each of the outcomes. A K-tuple of patients is well discriminated for a state k if the patient in state k was predicted to have the highest probability of being in state k compared to the others in the K-tuple. If we let patients i_j be a patient in state j, then the PDI for state k in that K-tuple can be given as:

$$C_K^k(i_1, i_2, ..., i_k, ..., i_K) = \begin{cases} 1 & P_{i_k}^k > \max\left(P_{i_j} \ : \ j \neq k\right) \\ 0 & P_{i_k}^k < \max\left(P_{i_j} \ : \ j \neq k\right) \\ \frac{1}{m} & P_{i_k}^k = \max\left(P_{i_j} \ : \ j \neq k\right), \ m = \left|\left\{j \ : \ P_{i_j}^k = P_{i_k}^k\right\}\right| \end{cases}$$

This definition also includes the caveat that if there are ties for the maximum predicted probability by assigning $^1/_m$ when that occurs, where m is the number of patient (including i_k tying for highest probability). We can see that if K=2, this collapses to C_2 defined previously for the two states involved.

For a K-tuple of patients, we also define their combined IPCW as the product of their individual IPCWs. This allows us to define a PDI for a K-tuple in a given state.

$$\text{PDI}_K^k(i_1, i_2, ..., i_K) = \left(\prod_{j=1}^{K} \omega_{i_j}\right) C_K^k(i_1, i_2, ..., i_K)$$

This allows us to define average weighted PDI for a K-tuple of patients as:

$$\text{PDI}_K(i_1, i_2, ..., i_K) = \frac{1}{K}\sum_{k=1}^{K} \text{PDI}_K^k(i_1, i_2, ..., i_K)$$

Or, we can summarise by finding the average PDI for a given state across the whole population:

$$\text{PDI}_K^k = \left(\frac{1}{\prod_{k=1}^{K} N_k}\right) \sum_{i_1 \in A_1} \sum_{i_2 \in A_2} ... \sum_{i_K \in A_K} \text{PDI}_K^k(i_1, i_2, ..., i_K)$$

These averages can be averaged again to get an overall measure of PDI:

$$\begin{aligned}\text{PDI}_K &= \frac{1}{K}\sum_{k=1}^{K} \text{PDI}_K^k \\ &= \left(\frac{1}{\prod_{k=1}^{K} N_k}\right) \sum_{i_1 \in A_1} \sum_{i_2 \in A_2} ... \sum_{i_K \in A_K} \text{PDI}_K(i_1, i_2, ..., i_K)\end{aligned}$$

Similar to the c-statistic, the P-model would score a PDI of 1, however the NI-model would achieve a PDI of $^1/_K$. Therefore, we need to adjust this PDI to correct the scaling to be that of he common c-statistic:

$$\hat{c} = (\text{PDI}_K)^{log_K(2)}$$

Since this new measure is on the same scale as the c-statistic, we can just refer to it as such.

As with the c-statistic, we can use an extended variant of DeLong's method for calculating the variance (and thus standard error and confidence intervals)

$$\text{var}(\hat{c}) = \sum_{k=1}^{K} \frac{1}{N_k} S_k$$

Where S_k are defined as:

$$S_k = \frac{1}{N_k - 1} \sum_{i_k \in A_k} (V_k(i_k) - \text{PDI}_K)^2$$

and $V_k(i_k)$ can be thought of as the PDI for an *individual* in a given state, i.e. the average of all PDI_k^k values that contain that individual

$$V_k(i_k) = \frac{1}{N_k} \sum_{i_1 \in A_1} ... \sum_{i_{k-1} \in A_{k-1}} \sum_{i_{k+1} \in A_{k+1}} ... \sum_{i_K \in A_K} \text{PDI}_K^k\,(i_1, ..., i_k, ..., i_K)$$

Computational Limitations

One major drawback of the PDI is that for large datasets and/or with many states, it can be computationally intensive. Therefore, an estimated PDI can be found by taking a sample of the K-tuples. To ensure robustness against censoring, each K-tuple should be drawn into th sample with probability inverse to the IPCW of that sample, where the IPCW of a K-tuple is calculated above as the probability of its elements. This is equivalent to drawing patients from each outcome with probability ${}^{\omega_j}/_{N_\omega}$. In this case, the calculations of the PDI remain similar, but each patient would be reset with a $\omega_j = 1$ (as the weighting has already been applied during sampling). This would also allow for an empirical estimate of a confidence interval.

4.4.4 Calibration - Multinomial Intercept, Matched and Unmatched Slopes

Since the traditional calibration metrics described above use a binomial logistic regression, it seems logical that the multi-dimensional extension for a multi-state models uses a multinomial logistic regression to provide parallel interpretation [192]. Unlike the other measures, we must choose a state to be our base-state, $k = 1$, this is usually the most populous initial state; however this choice is arbitrary and clinical reasoning may lead to a more logical choice.It is important to note that the values of β found below will depend on the choice of base-state, e.g. $\beta_{3,k}$ will be different if we choose the first versus the second state (or any other states in the model). It is common to cycle through the different choices of base-state prior to check whether the conclusions are similar regardless of the choice. We then estimate the following series of regressions for all $k > 1$:

$$\mathrm{E}\left[\log\left(\frac{O^k}{O^1}\right)\right] = \alpha_k + \beta_{2,k}\log\left(\frac{P^2}{P^1}\right) + ... + \beta_{K,k}\log\left(\frac{P^K}{P^1}\right)$$

Once again, using ω_i as weights for each patient during the regression process. This process estimates the α and β to provide a $(K-1)$ length vector of intercept terms, $\hat{\alpha} = \{\hat{\alpha_2}, \hat{\alpha_3}, ..., \hat{\alpha_K}\}$ and a $(K-1) \times (K-1)$ dimension matrix of slope terms, $\hat{\beta}$ with subscripts running from 2 to K in both dimensions.

The baseline models produce values similar to those found in the traditional calibration intercept and slope metrics, but directly extended to a multi-dimensional space. The P-Level for the Intercept would therefore be the zero-vector of length $(K-1)$ and the Slope would be the Identity matrix for $(K-1)$ dimensions. The NI-Level for the Intercept would be the prevalence (without the first state), $\{Q_2, Q_3, ..., Q_K\}$, and the Slope would be the zero-matrix for $(K-1)$ dimensions.

Software packages that can produce multinomial logistic regression [202] can also automatically produce confidence intervals surrounding these estimates.

As discussed earlier, traditional calibration measures are often associated with an interpretation depending on whether the model over- or under-predicts or has predictions that are too extreme or too moderate. Because of this, the multinomial extensions of these metrics cannot be aggregated to a single value (as with the other performance metric extensions), since doing

so would lose a lot of information. Instead, we discuss their interpretations of different elements of the Intercept vector and Slope matrix

The Intercept vector can be interpreted to have an element for every state except the default one. Therefore in our example, the first element of the Intercept vector is associated with the RRT state (state 2) and the second element is the Death state (state 3). Similarly, the Slope matrix has its rows and columns associated with these states (in the same order).

Intercept values below 0 imply that the associated state is over-predicted by the model, and values above 0 imply the state is under-predicted. By summing the entire Intercept vector, we can get a feel for how well the default state is calibrated. If the sum is below 0 (implying that on aggregate all other states are over-predicted) it implies that the default state is under-predicted, and vice versa.

The Matched-Slope, or the diagonal of the Slope matrix can be thought of as a vector with a single state associated with it (since the row-state and the column-state are the same state). If the values in the Matched-Slope are below 1, it implies that the predictions for that state are too moderate and if they are above 1, it implies that they are too extreme.

The Unmatched-Slope allows us to assess the Assumption of Independence of Irrelevant Alternatives (IIA) [203]. This assumes the predictions of one state is removed, the ratio of the observations in the other states will stay the same. More specifically, when dealing with a row/column state pair it means that we have found a correlation between the predictions of the row-state and the observations of the column-state. This implies that if the row-state were removed from our model and we were to re-standardise the other predictions, the predictions from the column-state would differ *more* than would be expected if we only relied on the information from the Matched-Slope.

If the Unmatched-Slope for the row/column-state pair is less than 0, then the removal of the row-state implies that the predictions from the new model would more *over-predict* the column-state compared to the original model. For complicated models, this kind of measure can provide insight into which states could potentially be dropped if they have a strong effect on other state's predictions. This is especially true if these changes to the predictions could potentially counteract the over/under-prediction found in the Matched-Slope.

4.5 Application to Real-World Data

4.5.1 Accuracy

Due to the prevalence of the different states in our population, the NI-levels for each the trivial extensions, and indeed the Multi-State version of the Brier Score would be different. These NI-levels can be seen 4.4, and in order for the Brier Score to be considered better than Non-Informative, it would have to be *lower* than these values. Fortunately, due to the correction applied to the aBS, the NI-level and P-level are 0 and 1 respectively, regardless of population.

Amongst the Pairwise, One-Vs-All and Transition based Brier Scores, the lowest (best) score is the RRT vs All score of 0.025, which translates to an adjusted Brier Score of 0.774. However, the highest adjusted Brier Score is 0.818 for the CKD vs All transition. The MSM has an aBS of 0.802 indicating a very strong discrimination level.

Table 4.4: NI-level for the different populations Brier Scores

One Vs All		
CKD vs ALL	Death vs All	RRT vs All
0.245	0.167	0.056
Pairwise		
CKD vs Death	CKD vs RRT	RRT vs Death
0.197	0.086	0.171
Trainsition		
CKD to Death	CKD to RRT	RRT to Death
0.185	0.124	0.211
MSM		
0.472		

Table 4.5: Measures of Accuracy for the Trivial extensions and Multi-State Model method with 95% Confidence Intervals shown

One Vs All			
	CKD vs All	Death vs All	RRT vs All
Brier Score	0.089 (0.077, 0.101)	0.073 (0.059, 0.086)	0.025 (0.022, 0.029)
aBS	0.818 (0.844, 0.793)	0.782 (0.822, 0.742)	0.774 (0.805, 0.742)
Pairwise			
	CKD vs Death	CKD vs RRT	RRT vs Death
Brier Score	0.133 (0.120, 0.146)	0.034 (0.028, 0.040)	0.075 (0.065, 0.086)
aBS	0.662 (0.695, 0.630)	0.801 (0.834, 0.767)	0.780 (0.811, 0.750)
Transition			
	CKD to Death	CKD to RRT	RRT to Death
Brier Score	0.093 (0.078, 0.107)	0.053 (0.045, 0.061)	0.095 (0.085, 0.104)
aBS	0.750 (0.790, 0.711)	0.785 (0.817, 0.752)	0.776 (0.798, 0.753)
MSM			
Brier Score	0.094 (0.067, 0.120)		
aBS	0.802 (0.858, 0.745)		

Table 4.6: Measures of Discrimination for the Trivial extensions and Multi-State Model method with 95% Confidence Intervals shown

One Vs All			
	CKD vs All	Death vs All	RRT vs All
c-statistic	0.689 (0.677, 0.701)	0.673 (0.659, 0.686)	0.625 (0.622, 0.629)
Pairwise			
	CKD vs Death	CKD vs RRT	RRT vs Death
c-statistic	0.733 (0.720, 0.746)	0.704 (0.688, 0.740)	0.675 (0.665, 0.686)
Transition			
	CKD to Death	CKD to RRT	RRT to Death
c-statistic	0.693 (0.678, 0.707)	0.701 (0.683, 0.730)	0.695 (0.685, 0.704)
MSM			
c-statistic	0.702 (0.658, 0.745)		
PDI	0.614 (0.595, 0.636)		

The two sets of Confidence Intervals are roughly consistent, which demonstrates that the population-based values are suitable for use when estimating the standard errors for these metrics.

4.5.2 Discrimination

For the c-statistic, the NI-level is 0.5 and P-level is 1, however for the PDI in the three state model, the NI-level is 0.333 and P-level is 1, which is why we adjust the PDI to coincide with the c-statistic. The CKD vs Death comparison achieves the highest c-statistic with 0.733, however the MSM score is still significant with a PDI of 0.614 against an NI-level of 0.333, which converts to a c-statistic of 0.702. The calculated confidence interval for the PDI (and thus the overall MSM c-statistic) are very narrow. This would be due to the large number of comparisons being made (the product of the population of all states involved).

4.5.3 Calibration

The P-level for the Intercept is 0, or a vector of 0's (of the same length as the number of states). Several of the estimates for the intercept were significantly different from 0 at the 5% level (i.e. their confidence interval did not include 0). Death vs All was significantly above 0 and RRT vs All was significantly below 0, indicating under- and over-predictive behaviours, respectively. RRT vs Death was not significantly different from 0, but the other two Pairwise values were under-predictive, and all of the Transition measures of the c-statistic were over-predictive. In the MSM results, both the Intercept values were statistically below 0 meaning that our predictions were over-predicting both of these states, which indicates that the CKD state was being under-predicted by our model.

For the Calibration Slope, we would ideally have a P-level of 1 for the traditional Slope, and a 2x2 identity matrix for the MSM extension. The Slope for the RRT vs All group was slightly above 1 indicating that the model's predictions are grouped too tightly and should be spread out within the prediction range.

For the CKD vs Death Pairwise comparison, the Slope of 0.881 is statistically significantly

Table 4.7: Measures of Intercept for the Trivial extensions and Multi-State Model method with 95% Confidence Intervals shown

One Vs All			
	CKD vs All	Death vs All	RRT vs All
Intercept	0.004 (-0.011, 0.019)	0.048 (0.023, 0.073)	-0.062 (-0.079,-0.045)
Slope	0.993 (0.957, 1.028)	1.011 (0.986, 1.035)	1.034 (1.013, 1.056)
Pairwise			
	CKD vs Death	CKD vs RRT	RRT vs Death
Intercept	0.051 (0.030, 0.071)	0.029 (0.015, 0.043)	-0.032 (-0.033,-0.030)
Slope	0.881 (0.863, 0.898)	1.013 (1.008, 1.018)	0.953 (0.918, 0.988)
Transition			
	CKD to Death	CKD to RRT	RRT to Death
Intercept	-0.029 (-0.057,-0.001)	-0.040 (-0.065,-0.016)	-0.032 (-0.033,-0.030)
Slope	1.018 (0.974, 1.062)	0.940 (0.924, 0.957)	0.958 (0.942, 0.975)
MSM			
Intercept	$\begin{bmatrix}-0.033\\-0.036\end{bmatrix}\left(\begin{bmatrix}-0.061\\-0.048\end{bmatrix},\begin{bmatrix}-0.005\\-0.024\end{bmatrix}\right)$		
Slope	$\begin{bmatrix}1.110 & -0.073\\-0.046 & 1.118\end{bmatrix}\left(\begin{bmatrix}1.086 & -0.085\\-0.062 & 1.113\end{bmatrix},\begin{bmatrix}1.134 & -0.061\\-0.030 & 1.123\end{bmatrix}\right)$		

lower than 0 and indicates that the predictions are too sparse and would need to be gathered inwards, this is reflected in figure 4.2 where the CKD vs Death curve in the middle plot is visibly different to the diagonal dotted line, which represents perfect calibration.

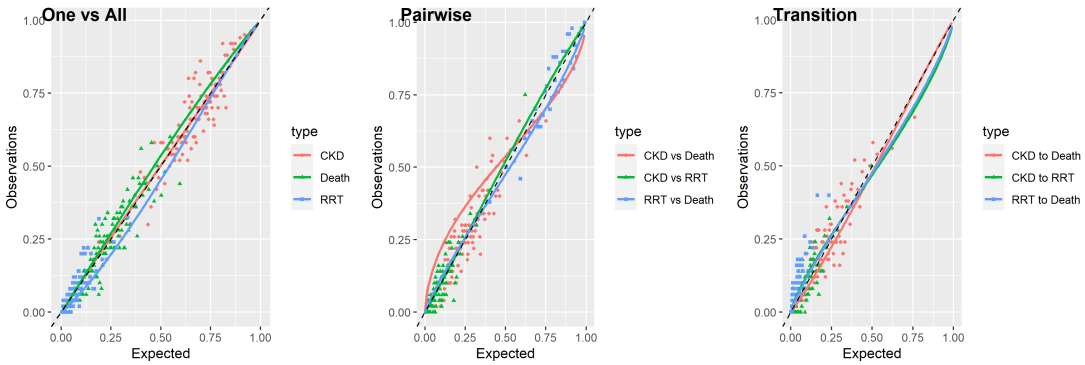

Figure 4.2: Calibration plot for the three trivial extensions, One Vs All, Pairwise and Transition

For the 2x2 matrix produced for the MSM extension, the Matched-Slope is statistically above 1 for the Death state (the second value on the diagonal 2), this indicates that the predictions of the Death state are slightly too moderate to match the observed data, and would need to be spread slightly to match. The Unmatched-Slope (those taken from the off-diagonal) are all significantly below 0, meaning that the model does not strictly follow the independence of irrelevant alternatives assumption and therefore the removal of one of the states *does* have a small effect on the others.

4.6 Discussion

Model Validation is an important step in assessing how generalisable and transportable a clinical prediction model will be to a new population. In this paper, we have extended the current methods of model validation to a Multi-State framework. The extensions have included adapting methods that were previously focused on single outcome cross-sectional models to incorporate multiple outcome states and account for censoring. R code to calculate the metrics is available in Appendix D. These extensions and their relevant adjustments have been normalised to allow for their predictive ability to be understandable at the same scale regardless of the number of states and prevalence within a population. This can allow researchers to directly compare the predictive ability of more complicated models to simpler ones.

The results from applying our validation metrics to our model show consistency with trivial extensions (where traditional metrics are applied to sub-sections of the model), which adds to their robustness.

For the Brier Score, the extension is a natural one, relying on the original multinomial definition and also provides an updated version, the aBS, which sets the level of a Non-informative (also known as a Random Model) to be 0 with a Perfect model scoring 1. Models that are worse than a Non-Informative model, score less than 0. Although not useful for direct application, models that score in this negative range can be used to inform other models or by simply altering their outcomes (e.g. by swapping states), a new model can be developed (which would have to undergo a similar validation process). This standardisation can also be useful in non-MSM situations, when there is a need to compare populations with different prevalences. The results from our model demonstrate consistency between the extensions to the traditional

The Discrimination, which is traditionally most commonly measured using the c-statistic has various ways of being extended, most of which have been studied by Van Calster et al [193]. We chose to use their Polytomous Discriminatory Index, as it was the most robust one and provided a reasonable level for a Non-Informative model. It was also simple to standardise to the same scale as the more well-known c-statistic.

The Calibration measurements provided more complexity in their extensions as the concept of a Calibration Intercept and Calibration Slope have been well studied elsewhere [192], [204], the effect of inter-state dependency has not. The Intercept provides an understanding of how over- or under-predicted a state is within a population, and thus also provides an idea of how under- or over- predicted the "default" state is. The Matched-Slope (i.e. the diagonals of the Slope matrix) is interpretable similar to the standard Calibration Slope as it provides an idea of how moderate predictions are compared to the observations and could be used to adjust a model if required.

The Unmatched-Slope provides an indication of how well the model satisfies the assumption of irrelevant alternatives (IIA). If the IIA had held for the Unmatched-Slope, it would imply that the removal of the one state, for example, RRT, would simply normalise the other predictions (of the CKD and Death states) to have the same ratio, but so that they sum to 1. However, the value in the RRT row and Death column (top right of our results) is negative and statistically different from 0. The breaking of this assumption implies that the removal of the RRT state would *increase* the values of the predictions in the Death state compared to those in the CKD

state, therefore:

$$\frac{\mathrm{E}\left[\mathrm{P}\left(\mathrm{Death}|\ !\mathrm{RRT}\right)\right]}{\mathrm{E}\left[\mathrm{P}\left(\mathrm{CKD}|\ !\mathrm{RRT}\right)\right]} > \frac{\mathrm{O}\left[\mathrm{P}\left(\mathrm{Death}\right)\right]}{\mathrm{O}\left[\mathrm{P}\left(\mathrm{CKD}\right)\right]}$$

The value in the bottom-left of the Calibration Slope Matrix is also negative and so the same interpretation happens for the values of the RRT predictions if Death were removed as a state. Since these two values are statistically close, however, this implies that the removal of CKD would have little effect on the ratio of the predictions for RRT with Death.

One limitation of the extensions provided is that they are applied at individual snapshots of time, and thus would need to be calculated as such. Future work could involve adapting these methods to be applicable towards a smoothed curve over time. For example, currently the PDI would be recalculated for the entire population at a given time-point, a smoothed measure would "update" the PDI when an event has occurred as populations move through time. It would also be important to adapt multi-nomial methods of re-calibration, such as those developed by Hoorde et al [192], [204], to a multi-state framework as described here.

The metrics extended herein do not encompass the entire literature of available metrics which had the potential to be extended into the MSM field. The three metrics used were chosen partly due to popularity within the clinical literature and how easy they are to interpret (which are incidentally not unrelated measures themselves). For example, to measure the discrimination of a model, we could have chosen to extend the D-statistic [69] through methods similar to the above for the c-statistic, however the interpretation of such a metric and what is considered 'good' would be difficult to establish (whereas this has been studied prior for the PDI). Alternatively, we could have extended the psuedo-observation method (as discussed in Chapter 3) to assess model calibration. However this would be computationally much more intensive than the logistic regression method and is intrinsically more complicated to implement. The results from such a model would be interpreted similar to those of a logistic regression based method and so could be cause for further study. The authors highly encourage the extension and establishment of other metrics to measure these (and other factors associated with model quality), as other metrics may be more appropriate depending on the model being assessed and the clinical utility of such a model.

Although some of the methods demonstrated here were developed by others in categorical outcome data [191], [193]; we are the first to apply them to a Multi-State scenario. This included the application of the IPCW to account for a time-trend and censoring and providing suitable adjustments to allow for cross-comparisons, regardless of the number of states. Before this work, it was previously un-assessable whether the additional information (e.g. prediction of Death alone or Death from multiple causes) came at a cost to model performance.

Chapter 5

Development and External Validation of a Multi-State Clinical Prediction Model for Chronic Kidney Disease Patients Progressing onto Renal Replacement Therapy and Death

MA Barrowman, GP Martin, N Peek, M Lambie, W Hulme, R Chinnadurai, J Lees, P Kalra, P Mark, J Traynor, M Sperrin

Abstract

Background

Clinical Prediction Models (CPMs) provide individualised predictions for patient outcomes. Traditionally, these models provide predictions for single outcomes, however in many circumstances, the ability to predict a multi-dimensional outcome with a single model can be advantageous. Many CPMs have been developed to predict the risk of different outcomes in individuals following chronic kidney disease (CKD) onset, but few allow the ability to predict the risk of patients transitioning onto renal replacement therapy (RRT) as well as death. For example, the risk of having a transplant within 1 year following dialysis, or the risk of remaining on dialysis until death. Multi-state models provide the vehicle to make such predictions, but have not been used within the CKD context.

Objective

Our objective was to develop a Multi-State Clinical Prediction Model (MSCPM), which can be used to predict patient progression through three states, untreated CKD (CKD), Renal Replacement Therapy (RRT) and Death (Dead).

Methods

We developed our model using tertiary care data from the Salford Kidney Study (SKS) as our development data set and secondary care data from the West of Scotland (SERPR) dataset as our external validation set. State transition were modelled using the Royston-Parmer regression technique and combined to create a single model. Model performance was assessed for accuracy, discrimination and calibration using methods both internally and externally. The model was then used to create an online calculator.

Results

Age was a strong predictor of mortality and outcomes were highly dependent on primary renal diagnosis. Models performed well in both the internal and external validation with a Brier Score of 0.67/0.62 (internal/external, respectively), c-statistic of 0.83/0.81 and an averaged calibration intercept of 0.00/0.00 and slope diagonal of 1.34/1.53 (indicating under-prediction of all non-untreated CKD states for more extreme values).

Discussion

Our CPM provides clinicians and patients with multi-dimensional predictions across different outcome states and any time point. This implies that users of these models can get more information about their potential future without a loss to the model's calibration nor its discriminative ability.

Supplementary Material

Supplementary Material is available in Appendix F.

5.1 Introduction

A clinical prediction model (CPM) is a tool which provides patients and clinicians with a measure of how likely a patient is to suffer a specific clinical event, more specifically, a prognostic model allows the prediction of future events [3]. CPMs use data from previous patients to estimate the outcomes of an individual patient. Prognostic models can be used in clinical practice to influence treatment decisions.

Within Chronic Kidney Disease (CKD), prognostic models have been developed to predict mortality [205]–[206], End-Stage Renal Disease [207], the commencements of Renal Replacement Therapy (RRT) [208], [209]–[210] or mortality after beginning dialysis [211]–[212]. Some previous models have used the commencement of RRT as a proxy for CKD Stage V [213]–[214], while

others have investigated the occurrence of cardiovascular events within CKD patients[215]–[216]. Reviews by Grams & Coresh [217], Tangri et al [218] and Ramspek et al [219], which explored the different aspects of assessing risk amongst CKD or RRT patients, found that the current landscape of CKD prediction models is lacking from both a methodological and clinical perspective [13], [85].

Methodologically, the majority of existing CKD prediction models fail to account for competing events [220], [206], [221], have high risks of bias [205], [207], [209] or are otherwise flawed compared to modern clinical prediction standards [3], [13]. These can have large implications if these models were to be used in clinical practice as, for example, patients could be given a predicted probability of RRT, which has not been adjusted for the probability of death and is thus, in fact a probability of RRT, assuming you don't die (a very different value if you have a high risk of death).

In 2013, Begun et al [141] developed a multi-State model for assessing population-level progression through the severity stages of CKD (III-V), RRT and/or death, which can be used to provide a broad statement regarding a patient's future. In 2014, Allen et al [222] applied a similar model to liver transplant recipients and their progression through the stages of CKD with a focus on the predictions of measured vs estimated glomerular filtration rate (mGFR vs eGFR). In 2017, Kulkarni et al [210] developed an MSM focusing on the categories of Calculated Panel Reactive Antibodies and kidney transplant and/or death.

Most recently, in 2018, Grams et al [140] developed a multinomial clinical prediction model for CKD patients which focused on the occurrence of RRT and/or cardiovascular events. As of the publication of this paper, this is the only currently existing CPMs of this kind for CKD patients.

However, the first three of these existing models (Begun, Allen and Kulkarni) categorise continuous variables to define their states at specific cut-offs and this has been shown to be inefficient when modelling [30] and none of these models have undergone any validation process, whether internal or external [11].

It is also important to note that although these models can be used to predict patient outcomes, they were not designed to produce individualised patient predictions as is a key aspect of a clinical prediction model; they were designed to assess the methodological advantages of MSMs in this medical field, to describe the prevalence of over time of different CKD stages and to produce population level predictions for patients with different levels of panel-reactive antibodies [10].

The fourth model (Grams), is presented as a Multi-State Model and the transitions involved were studied and defined, however the underlying statistical model is two multinomial logistic models analysed at 2 and 4 years, which assumes homogeneity of transition times. Two downsides to the implementation of this model are that it can only produce predictions at those predefined time points and that it is unable to estimate duration of time on dialysis.

Therefore, our aim is to develop a MSCPM - we do this by modelling patient pathways through a Multi-State Model by choosing transition points which can be exactly identified and include states which produce a clinical difference in patient characteristics. Our modelling techniques allow for individual predictions of multi-dimensional outcomes at any time point .

The models produced by this process will then be validated, both internally and externally, to compare their results and demonstrate the transportability of the clinical prediction models.

5.2 Methods

We report our work in line with the TRIPOD guidelines for development and validation of clinical prediction models [13], [14].

5.2.1 Data Sources

The models were developed using data from the Salford Kidney Study (SKS) cohort of patients (previously named the CRISIS cohort), established in the Department of Renal Medicine, Salford Royal NHS Foundation Trust (SRFT). The SKS is a large longitudinal CKD cohort recruiting CKD patients since 2002. This cohort collects detailed annualised phenotypic and laboratory data, and plasma, serum and whole blood stored at -80°C for biomarker and genotypic analyses. Recruitment of patients into SKS has been described in multiple previous studies [223], [224] and these have included a CKD progression prognostic factor study and to evidence the increased risk of cardiovascular events in diabetic kidney patients. In brief, any patient referred to Salford renal service (catchment population 1.5 million) who is 18 years or over and has an eGFR measurement of less than 60ml/min/1.73m^2 (calculated using the CKD-EPI formula [225]) was approached to be consented for the study participation.

At baseline, the data, including demographics, comorbidities, physical parameters, lab results and primary renal diagnosis are recorded in the database. Patients undergo an annual study visit and any changes to these parameters are captured. All data except blood results are collected via questionnaire by a dedicated team of research nurses. Blood results (baseline and annualised), first RRT modality and mortality outcome data are directly transferred to the database from Salford's Integrated Record [226]. eGFR, uPCR, comorbidity and blood results were measured longitudinally throughout a patient's time within the cohort.

Due to limitations in our data, we were agnostic to how long since patients were diagnosed with CKD. Therefore, we defined a patient's start date for our model as their first date after consent at which their eGFR was recorded to be below 60ml/min/1.73m^2. Some patients consented with an eGFR that was already below 60, and some entered our study later when their eGFR was measured to be below 60. This implies that our models includes both patient who have recently been diagnosed with CKD eGFR $\lessapprox 60$ *and* those that have been suffering with CKD for an arbitrary amount of time. This timelessness of the model means it can be applied to any patient prior to commencement of RRT at any time after their initial CKD diagnosis, be that immediately or 10 years later.

All patients registered in the database between October 2002 and December 2016 with available data were included in this study. As this is a retrospective convenience sample, no sample size calculations were performed prior to recruitment. However, we were able to use the given sample size to calculate the maximum number of predictors for each transition as per Riley, et al [46], which permitted us to use 30 predictors for each transition in the Three-State Model, although permitted number of variables were significantly lower in the Five-State

Model and thus this model would be susceptible to overfitting. All patients were followed-up within SKS until the end-points of RRT, death or loss to follow-up or were censored at their last interaction with the healthcare system prior to December 2017. Date of death for patients who commenced RRT was also included in the SKS database.

For external validation of the model, we extracted an independent cohort from the West of Scotland Electronic Renal Patient Record (SERPR). Our extract of SERPR contains all patients known to the Glasgow and Forth Valley renal service who had an eGFR measure of less than 60ml/min/1.73m^2 between January 2006 and January 2016. This cohort has been previously used in Chronic Kidney Disease Prognosis consortium studies investigating outcomes in patients with CKD [227] and a similar cohort has been used for the analysis of skin tumours amongst renal transplant patients. Use of anonymised data from this database has been approved by the West of Scotland Ethics Committee for use of NHS Greater Glasgow and Clyde 'Safe Haven' data for research.

Both the internal and external validation cohort were used as part of the multinational validation cohort used by Grams et al in their multinomial CPM discussed above [140]. In SERPR, start dates were calculated to be the first time point where the following conditions were met:

- eGFR is measured at less than 60
- There is at least one prior eGFR measurement
- Patient is 18 or over

The second requirement was implemented to avoid a bias in the eGFR Rate. eGFR Rate is a measure of the change in eGFR over time and is calculated as the difference between the most recent two eGFR measurements divided by the time between them. For patients who entered the system with an eGFR < 60, their eGFR Rate would be unavailable (i.e. missing). Otherwise, patient eGFRs would *have* to drop to below 60 and thus eGFR Rate would be negative. In addition, to avoid Acute Kidney Incident events, we also filtered measurements of eGFR Rate that were more extreme than 1.5x those found in the SKS population.

5.2.2 Model Design

Three separate models were developed, so we could determine a clinically viable model while maintaining model parsimony as much as possible: a Two-State, Three-State and Five-State model, each building on the previous models' complexity (see figure 5.1). The Two-State model was a traditional survival analysis where a single event (death) is considered. The Three-State model expanded on this, by splitting the Alive state into transient states of (untreated) CKD and (first) RRT; patients can therefore transition from CKD to Death or CKD to RRT, and then onto RRT to Death. The Five-State model stratifies the RRT state into HD, PD and Tx and allows similar transitions into and out of the RRT states; however, the transition from Tx to Death was not considered as it was anticipated a priori that there would be insufficient patients undergoing this transition and that the process of undergoing a transplant would be medically transformative and so it would be inappropriate to assume shared parameters

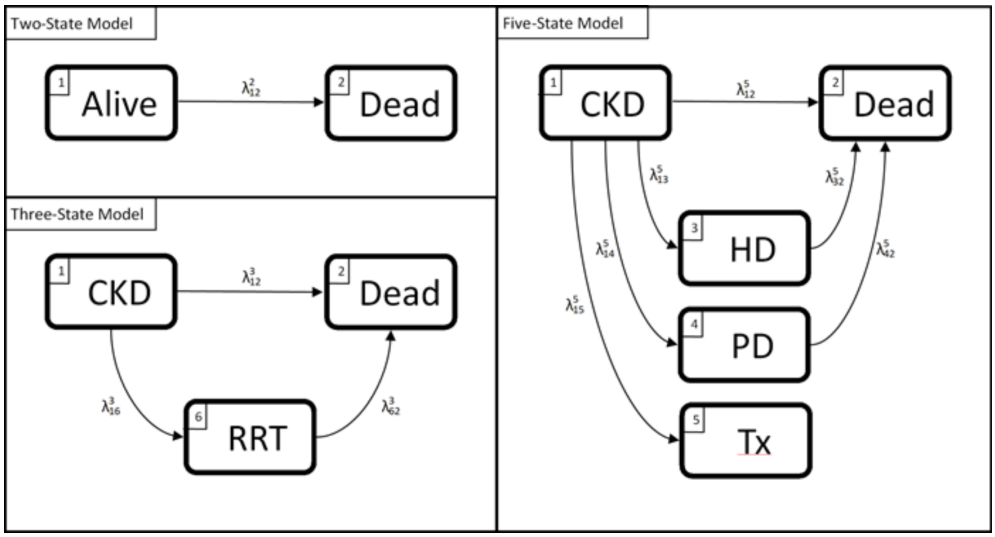

Figure 5.1: Diagram of the three models, the states being modelled and relevant transitions

before and after the transition (i.e. Tx was modelled as a second absorbing state). Missing data was handled using multiple imputation [75] using times for all events as imputation covariates. Variables considered as covariates were demographics (sex, age, smoking status and alcohol consumption), comorbidities (congestive cardiac failure (CCF), chronic obstructive pulmonary disease (COPD), prior cerebrovascular accident (CVA), hypertension (HT), diabetes mellitus (DM), ischemic heart disease (IHD), chronic liver disease (LD), prior myocardial infarction (MI), peripheral vascular disease (PVD) and slid tumour (ST)), physical parameters (BMI, blood pressure), blood results (haemoglobin, albumin, corrected calcium and phosphate measures), urine protein creatinine ratio (uPCR) and primary renal diagnosis (grouped as per ERA-EDTA classifications [228]). From these variables, uPCR and eGFR Rate of change were also derived [229], [230] as well as their log transforms and log(Age) and Age^2. For each transition, Backwards-forwards stepwise methods were used to minimise the AIC, allowing different variables to be used for each transition. Variables were also assessed for their adherence to the proportional hazards assumption by calculating their Schoenfeld residual [231], [232], for those that fail the test of proportionality, we include a time trend in their hazard estimate.

5.2.3 Other Considerations

Data was recorded in a time-updated manner, however all variables were measured at baseline to emulate the real-world application of the model (i.e. future prediction of states and not covariates). Race was also assessed in the populations, but due to high heterogeneity, as most patients were white, it was omitted as a potential predictor from the models.

Intermediate states (RRT or modality) were considered to be medically transformative, and so a semi-markov (clock reset) method for analysis was considered to be well justified [233]. Each transition was modelled under a proportional hazards assumption using the Royston-Parmar technique [67] to estimate coefficients for each covariate and a restricted cubic spline (on the log-time scale) for the baseline cumulative hazard, see Section E.1. The hazards for

each transition can be combined to produce estimates for the probability of a patient being in any state at any time [113], see Section E.2.

All missing data were assumed to be missing at random and so were multiply imputed using chained equations with the Nelson-Aalen estimators for each relevant transition as predictors [75] in the MICE model. Some variables (smoking status and histories of COPD, LD and ST) were present in the SKS (development) dataset, but were completely missing in the SERPR extract (validation) and so these were multiply imputed from the development dataset [234] by attaching each of the imputed SERPR dataset to the SKS dataset and re-imputing once for each of the imputed datasets.

For variable selection, we stacked the imputed datasets together to create a larger, pseudo-population [235] and performed backwards-forwards selection based on minimising the AIC at each step. We included the γ time trends for each covariate in these steps if the covariate was also present. This was repeated for each transition and for different numbers of evenly spaced knots in modelling the form of the cubic spline hazard, m={0,1,2,3,4,5}. This allowed for different transitions to use different sets of variables and numbers of knots in the final model. Some combinations of variables resulted in models that were intractable and so these models were excluded. Once a set of variables were chosen, the R-P model was applied to each imputed dataset individually and the resulting coefficients and cubic spline parameters were aggregated across imputations using Rubin's Rules [236]. This gave a model fully defined by smooth cubic splines representing the individualised cause-specific cumulative hazard for each transition, which can be smoothly derived to estimate an individuals cause-specific hazard function.

5.2.4 Validation

Each of the three models were internally validated in the development dataset using bootstrapping to adjust for optimism and then further externally validated in the validation dataset extracted from SERPR [237]. The bootstrapping method was also used for both validations to adjust the results for optimism and to produce confidence intervals around the performance metric estimates. To assess the performance in low eGFR patients, the models were also validated in subsets of the SKS and SERPR where patients had an eGFR $< 30/\text{ml}/\text{min}/1.73\text{m}^2$.

Model accuracy was assessed using the Brier Score, discrimination was assessed using the c-statistic and the calibration was assessed using the multi-dimensional intercept and slope-matrix, as described in Chapter 4. These measures were taken at One-Year, Two-Years and Five-Years after the patient's start dates, and therefore are equivalent to validation measures at that point in time after a prediction has been applied.

Further details of how the models were developed and validated is discussed in the Supplementary materials in appendix F.

5.2.5 Example

Once the models have been developed, we will apply them to three example patients to demonstrate their use and applicability to the general population. We will provide a direct clinical

Table 5.1: Details of the Example Patients

Vars	Patient 1	Patient 2	Patient 3
Age	20	40	66
Gender	Female	Male	Female
Smoking Status	Non-Smoker	Smoker	Non-Smoker
BP	144/101	160/90	140/80
Albumin	39	40	40
Correct Calcium	2.3	3.0	2.6
Haemoglobin	150	100	14
Phosphate	0.68	2.00	0.86
eGFR	42	10	51
eGFR Previous	50 (one week ago)	30 (one year ago)	70 (one week ago)
uPCR	0.30	0.20	0.01
uPCR Previous	0.80 (one month ago)	1.20 (one year ago)	0.06 (one week ago)
Primary Diagnosis	Glomerulonephritis	Tubular Necrosis	Diabetes
Comorbities	COPD[a] LD[b] ST[c]		DM[d] COPD[a] HT[e]

[a] Chronic obstructive pulmonary disease
[b] Liver Disease
[c] Solid Tumour
[d] Diabetes Mellitus
[e] Hypertension

estimation of these patient outcomes based on years of nephrological experience and compare this with the results presented by our clinical prediction model.

We have chosen three (synthetic) patients to use as examples of the use of our model. Their details can be seen in table 5.1. Our three example patients cover a broad range of ages and other covariates. A clinically guided prediction for these patients would assume that Patient 1 has a high chance of proceeding as normal (with little need for RRT), Patient 2 would be recommended to start RRT soon and Patient 3 would be predicted to have a high risk of mortality with or without RRT.

5.2.6 Calculator

As part of this work, we have also produced an online calculator to allow patients and clinicians to easily estimate outcomes without worrying about the mathematics involved.

All analysis was done in `R 3.6.2` [183] using the various `tidyverse` packages [184], as well as the `mice` [238], `flexsurv` [239], `nnet` [202] and `furrr` [240] packages. The calculator was produced using the `shiny` package [188].

5.3 Results

5.3.1 Data Sources

There were 2,981 patients in the SKS dataset and 7,709 patients in the SERPR dataset. As seen in table 5.2, the Age of the populations had a mean of 64.4 and 65.9 respectively with a very broad range. Due to the inclusion criteria, eGFR were capped at a maximum of 60, and was consistent across populations; however, the rate of change for eGFR was much wider in the SERPR patients than in the SKS, and it was decreasing much faster, on average (-25 vs 0) . The uPCR measures are presented in our results as g/mmol, rather than the more conventional g/mol, this is to better present results and coefficients of varying magnitudes. Levels of missingness were much higher in the SERPR dataset in most continuous variables.

Table 5.2: Population demographics, continuous displayed as median (Inter-Quartile Range), and Categorical/Comorbidity data as number (percent) range and number missing are also included

	SKS			SERPR		
Variable	median (IQR) or n (%)	range	missing (%)	median (IQR) or n (%)	range	missing (%)
Age	67.000 (19.000)	[20.000, 94.000]	0 (0.00%)	69.000 (17.000)	[11.000, 98.000]	0 (0.00%)
eGFR	28.612 (22.387)	[3.578, 59.966]	0 (0.00%)	35.398 (20.565)	[1.199, 59.994]	0 (0.00%)
eGFR Rate	-0.037 (0.294)	[-19.107, 33.782]	1278 (42.87%)	-0.932 (21.897)	[-28.636, 50.653]	0 (0.00%)
SBP	139.000 (29.000)	[77.000, 220.000]	50 (1.68%)	144.000 (30.000)	[82.000, 258.000]	6345 (82.31%)
DBP	75.000 (14.000)	[36.000, 159.000]	52 (1.74%)	77.000 (17.000)	[35.000, 128.000]	6345 (82.31%)
BMI	27.993 (7.842)	[13.182, 61.467]	572 (19.19%)	28.081 (6.811)	[17.073, 52.403]	7491 (97.17%)
Albumin	43.000 (5.000)	[12.000, 52.000]	60 (2.01%)	37.000 (6.000)	[7.000, 53.000]	3134 (40.65%)
Calcium	2.300 (0.180)	[1.210, 3.660]	68 (2.28%)	2.410 (0.160)	[1.455, 3.400]	4513 (58.54%)
Haemoglobin	122.000 (23.000)	[61.000, 195.000]	72 (2.42%)	116.000 (29.000)	[7.100, 208.000]	3557 (46.14%)
Phosphate	1.120 (0.320)	[0.430, 3.710]	87 (2.92%)	1.160 (0.320)	[0.320, 4.370]	4510 (58.50%)
uPCR	0.035 (0.103)	[0.001, 2.025]	245 (8.22%)	0.062 (0.177)	[0.001, 1.943]	7170 (93.01%)
uPCR Rate	-0.008 (0.188)	[-70.727, 28.199]	1777 (59.61%)	0.004 (0.905)	[-176.200, 952.812]	7495 (97.22%)
Gender:			0 (0.00%)			0 (0.00%)
Male	1865 (62.56%)			3885 (50.40%)		
Female	1116 (37.44%)			3824 (49.60%)		
Ethnicity:			0 (0.00%)			7009 (90.92%)
White	2875 (96.44%)			679 (8.81%)		
Asian	75 (2.52%)			12 (0.16%)		
Black	21 (0.70%)			7 (0.09%)		
Other	10 (0.34%)			2 (0.03%)		
Smoking Status:			42 (1.41%)			7709 (100.00%)
Former	1535 (51.49%)			0 (0.00%)		
Non-Smoker	979 (32.84%)			0 (0.00%)		
Smoker	379 (12.71%)			0 (0.00%)		
Former (More than 3Y)	46 (1.54%)			0 (0.00%)		
Diagnosis Group:			567 (19.02%)			6640 (86.13%)

Table 5.2: Population demographics, continuous displayed as median (Inter-Quartile Range), and Categorical/Comorbidity data as number (percent) range and number missing are also included *(continued)*

	SKS			SERPR		
Variable	median (IQR) or n (%)	range	missing (%)	median (IQR) or n (%)	range	missing (%)
Systemic diseases affecting the kidney	1304 (43.74%)			316 (4.10%)		
Glomerular disease	442 (14.83%)			278 (3.61%)		
Tubulointerstitial disease	268 (8.99%)			173 (2.24%)		
Miscellaneous renal disorders	227 (7.61%)			198 (2.57%)		
Familial / hereditary nephropathies	173 (5.80%)			104 (1.35%)		
Comorbidities:						
CCF	2414 (81.09%)		4 (0.13%)	408 (5.29%)		0 (0.00%)
COPD	2411 (80.99%)		4 (0.13%)	0 (0.00%)		7709 (100.00%)
CVA	2727 (91.60%)		4 (0.13%)	186 (2.41%)		0 (0.00%)
DM	992 (33.32%)		4 (0.13%)	1535 (19.91%)		0 (0.00%)
HT	2546 (91.48%)		198 (6.64%)	3114 (40.39%)		0 (0.00%)
IHD	2393 (80.38%)		4 (0.13%)	863 (11.19%)		0 (0.00%)
LD	2891 (97.11%)		4 (0.13%)	0 (0.00%)		7709 (100.00%)
MI	2492 (83.71%)		4 (0.13%)	556 (7.21%)		0 (0.00%)
PVA	2485 (83.47%)		4 (0.13%)	376 (4.88%)		0 (0.00%)
ST	2570 (86.33%)		4 (0.13%)	0 (0.00%)		7709 (100.00%)

Table 5.3: Event times for the two populations presented as Number of Events, Median, Inter-Quartile Range and Max

Transition			SKS				SERPR			
Model	From	To	n	median	IQR	max	n	median	IQR	max
Two	Alive	Dead	1,427	4.0 y	4.3 y	15.0 y	3,010	4.8 y	3.3 y	10.1 y
Three	CKD	Dead	1,125	3.5 y	4.2 y	15.0 y	2,568	4.7 y	3.3 y	10.1 y
		RRT	680	2.5 y	3.4 y	14.2 y	1,125	3.8 y	3.9 y	10.1 y
	RRT	Dead	302	2.2 y	3.3 y	13.5 y	442	1.6 y	2.5 y	9.2 y
Five	CKD	Dead	1,125	3.5 y	4.2 y	15.0 y	2,568	4.7 y	3.3 y	10.1 y
		HD	344	2.6 y	3.5 y	14.2 y	882	3.7 y	3.8 y	10.1 y
		PD	229	2.0 y	2.9 y	12.9 y	149	3.5 y	4.1 y	9.5 y
		Tx	107	3.2 y	2.7 y	12.1 y	94	5.0 y	4.4 y	9.8 y
	HD	Dead	185	2.1 y	3.3 y	11.8 y	394	1.6 y	2.5 y	9.2 y
	PD	Dead	107	2.4 y	3.2 y	11.8 y	47	2.1 y	2.3 y	8.5 y

Table 5.2 also shows a breakdown of the categorical variables across the populations. In the development population, crude proportions of males were numerically higher than females whereas in the validation population the proportions are much more matched (62.6% male vs 50.4% male). Most patients were white in the SKS dataset, and ethnicity has extremely high missingness in SERPR, which also contributed to its omission from the model.

Overall, there were high levels of comorbidities within the SKS population, but these levels were much lower in the SERPR population.

The median date for the date of death was 3.9 years in the SKS population and 4.9 years in the SERPR population. The median date for transition to RRT was 2.2 years and 1.5 years (in SKS and SERPR respectively). In SKS, transitions to HD happened 6 months later than PD, and in SERPR it was 3.6 months. The Maximum follow-up time in SKS was 15.0 years and in SERPR it was 10.1 years. This information can be seen in table 5.3.

5.3.2 Development

Here we present the proportional hazards for the Three-State Model in 5.4. The full model description, including non-proportional hazards and baseline hazards can be found in the Supplementary Materials in appendix F. Older patients are predicted to be likely to transition to RRT. Increased rates of decline of eGFR were associated with the transition from CKD to RRT.

Female patients are predicted to be more likely to remain in the CKD state than Males, or to remain in the RRT state once there. Smokers were predicted as more likely than Non-/Former Smokers to undergo any transition, apart from CKD to Tx. Blood results had associations with all transitions in some way, and disease aetiology were strongly associated with the transitions giving a wide range of predictions.

5.3.3 Validation

Table 5.5 shows the results from the internal validation in the Three-State Model at different time points and as an average over all time points upto 10-years from start-date. Both the discrimination and overall accuracy performances were strong with c-statistics ranging from

Table 5.4: Porportional Hazards for each transition in the Three-State Model

Var	CKD to Dead	CKD to RRT	RRT to Dead
constant term			
const	-4.007 (-15.263, 7.250)	1.292 (-1.654, 4.237)	-5.927 (-9.851, -2.002)
Age			
Age	-0.016 (-0.067, 0.035)		0.064 (0.052, 0.075)
log(Age)		-2.064 (-2.370, -1.758)	
eGFR			
eGFR	-0.067 (-0.116, -0.019)	-0.552 (-0.640, -0.465)	0.011 (-0.001, 0.023)
log(eGFR Rate)	0.044 (-0.118, 0.207)	0.260 (-0.023, 0.542)	
uPCR			
uPCR		1.403 (0.530, 2.275)	
Measures			
Albumin	-0.188 (-0.332, -0.044)	-0.025 (-0.059, 0.008)	-0.038 (-0.061, -0.016)
Corrected Calcium	2.383 (-0.720, 5.486)		
DBP	0.005 (-0.001, 0.010)		
Haemoglobin	-0.013 (-0.018, -0.008)	-0.007 (-0.013, -0.000)	
Phosphate	1.856 (0.124, 3.587)	0.483 (-0.050, 1.016)	
SBP		0.008 (0.004, 0.012)	
Gender (vs Male)			
Female	-0.562 (-1.656, 0.532)	-0.277 (-0.444, -0.110)	
Smoking Status (vs Former)			
Former (3y+)	-3.016 (-8.708, 2.676)		-132.164 (-265.921, 31.593)
Non-Smoker	-0.611 (-1.851, 0.630)		-0.186 (-5.373, 5.001)
Smoker	0.331 (-1.164, 1.827)		-2.492 (-9.124, 4.139)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies	-1.246 (-5.193, 2.702)	0.011 (-0.498, 0.521)	-0.958 (-4.616, 2.700)
Glomerular disease	0.170 (-1.526, 1.866)	-0.097 (-0.595, 0.401)	0.336 (-1.809, 2.481)
Miscellaneous renal disorders	1.404 (-0.472, 3.280)	-0.796 (-1.702, 0.109)	1.668 (-1.167, 4.502)
Tubulointerstitial disease	0.271 (-2.109, 2.650)	-0.605 (-1.157, -0.053)	1.572 (-0.972, 4.116)
Comorbidity			
CCF	-1.758 (-2.844, -0.673)		-2.840 (-6.378, 0.698)
COPD	-0.289 (-0.431, -0.147)		
CVA	-0.072 (-0.253, 0.109)		
DM	0.135 (0.003, 0.267)		0.208 (-0.075, 0.490)
IHD	0.108 (-0.035, 0.251)		
MI	-1.385 (-2.510, -0.260)		
PVD	-0.929 (-2.048, 0.191)		
ST	-1.331 (-2.553, -0.110)		-0.312 (-0.636, 0.012)

Table 5.5: Internal Validation of the Three-State Model, results presented as Estimate (95% CI, where possible)

Measure	eGFR	One Year	Two Year	Five Year	Average
Brier	< 30	0.74 (0.74, 0.75)	0.68 (0.68, 0.69)	0.64 (0.64, 0.65)	0.67 (0.67, 0.68)
	< 60	0.75 (0.74, 0.75)	0.73 (0.73, 0.73)	0.68 (0.67, 0.68)	0.68 (0.67, 0.68)
c-statistic	< 30	0.79 (0.72, 0.85)	0.79 (0.79, 0.79)	0.78 (0.78, 0.78)	0.78 (0.71, 0.86)
	< 60	0.93 (0.93, 0.93)	0.83 (0.82, 0.83)	0.72 (0.72, 0.72)	0.76 (0.73, 0.78)
Intercept	< 30	-0.04 (-0.05,-0.04) -0.06 (-0.07,-0.06) -0.05 (-0.07,-0.03)	-0.04 (-0.06,-0.03) -0.04 (-0.06,-0.02) -0.05 (-0.06,-0.04)	-0.06 (-0.07,-0.06) -0.04 (-0.04,-0.04) -0.05 (-0.06,-0.04)	-0.04 (-0.05,-0.03) -0.05 (-0.07,-0.03) -0.04 (-0.05,-0.04)
	< 60	0.00 (0.00, 0.01) 0.00 (-0.02, 0.03) 0.02 (0.02, 0.02)	-0.00 (-0.00,-0.00) 0.02 (0.02, 0.02) -0.01 (-0.02,-0.00)	0.00 (-0.03, 0.04) -0.01 (-0.03, 0.02) -0.00 (-0.03, 0.02)	0.00 (0.00, 0.01) 0.02 (-0.00, 0.04) 0.02 (-0.03, 0.07)
Slope	< 30	1.15, 0.10, 0.02 0.04, 1.08, 0.02 0.03, 0.05, 1.11	1.04, 0.02, 0.05 -0.03, 1.10, 0.00 0.01, 0.03, 1.05	1.13, 0.03, 0.02 -0.06, 1.13, 0.05 -0.06, 0.03, 1.08	1.10,-0.02, 0.05 0.01, 1.07, 0.07 0.13,-0.02, 1.12
	< 60	0.95, 0.07, 0.01 0.05, 1.00,-0.00 -0.04,-0.02, 1.01	1.01, 0.05,-0.09 0.04, 0.97, 0.01 -0.09, 0.06, 0.97	0.91,-0.05,-0.01 0.03, 0.92,-0.06 0.09, 0.03, 0.99	1.02,-0.01, 0.06 -0.03, 0.97, 0.00 -0.04, 0.03, 0.92

$0.72 - 0.93$ in the overall (<60 eGFR) population and $0.78 - 0.79$ in the <30 eGFR sub-population and a Brier Score ranging from $0.68 - 0.75$ in the overall (<60 eGFR) population and $0.64 - 0.74$ in the <30 eGFR sub-population.

Table 5.6 shows the results from the external validation in the Three-State Model at different time points and as an average over all time points upto 8-years from start-date (as longer times were not available for all states). As with the internal validation, both the discrimination and overall accuracy performances were strong. In the overall (<60 eGFR) population, the c-statistics ranges from $0.72 - 0.77$ in the <30 eGFR sub-population it ranges from $0.64 - 0.73$. Meanwhile, the Brier Score ranges from $0.63 - 0.72$ in the overall (<60 eGFR) population and $0.61 - 0.70$ in the <30 eGFR sub-population.

The calibration intercept results for the overall (<60 eGFR) population in the internal validation (als in Table 5.5) are all very close to 0, as expected. However, values in the <30 eGFR sub-population were significantly below 0, indicating that, for the RRT, RRT to Death and CKD to Death states, the model appears to systematically under-predict results and over-predict the CKD state. Similarly, for the overall (<60 eGFR) population, the calibration slope are close to 1 for most values at early time points, but at the five-year time point, the metric for the RRT state is significantly below 1, indicating a possible loss of predictive power over time. The calibration slope for the <30 eGFR sub-population were all significantly above 1, inidicating that, for the RRT, RRT to Death and CKD to Death states, the spread of the expected values were more condensed than was observed.

5.3.4 Example

The example patients seen in Table 5.1 were passed through our Three-State prediction model and the results for all time-points are shown in figure 5.2. The prognosis for all three patients were very different. Patient 1 (20 year old) had a very high probability of survival, with only an 16% chance of mortality by year 10 and 0% chance of commencing RRT. Patient 2 (40 year old) was predicted almost 90% chance of starting RRT, and over 70% chance of dying overall (either with or without RR). Patient 3 (66 year old) had a fast acceleration towards high mortality,

Table 5.6: External Validation of the Three-State Model, results presented as Estimate (95% CI, where possible)

Measure	eGFR	One Year	Two Year	Five Year	Average
Brier	< 30	0.69 (0.68, 0.69)	0.70 (0.69, 0.70)	0.61 (0.60, 0.61)	0.62 (0.62, 0.63)
	< 60	0.68 (0.67, 0.68)	0.72 (0.71, 0.72)	0.65 (0.64, 0.65)	0.63 (0.62, 0.63)
c-statistic	< 30	0.64 (0.62, 0.65)	0.65 (0.61, 0.68)	0.73 (0.71, 0.74)	0.68 (0.67, 0.70)
	< 60	0.74 (0.73, 0.75)	0.77 (0.76, 0.78)	0.72 (0.71, 0.72)	0.75 (0.71, 0.79)
Intercept	< 30	-0.21 (-0.22,-0.20)	-0.20 (-0.23,-0.17)	-0.21 (-0.21,-0.20)	-0.19 (-0.19,-0.19)
		-0.21 (-0.22,-0.19)	-0.21 (-0.23,-0.19)	-0.21 (-0.21,-0.20)	-0.22 (-0.23,-0.21)
		-0.21 (-0.28,-0.14)	-0.21 (-0.25,-0.17)	-0.19 (-0.20,-0.19)	-0.18 (-0.26,-0.11)
	< 60	-0.11 (-0.11,-0.11)	-0.09 (-0.10,-0.09)	-0.10 (-0.11,-0.09)	-0.12 (-0.14,-0.09)
		-0.09 (-0.11,-0.07)	-0.11 (-0.15,-0.07)	-0.09 (-0.09,-0.09)	-0.09 (-0.10,-0.09)
		-0.13 (-0.14,-0.12)	-0.11 (-0.12,-0.11)	-0.09 (-0.16,-0.03)	-0.10 (-0.12,-0.08)
Slope	< 30	1.36, 0.02,-0.03	1.19,-0.03,-0.05	1.28,-0.01,-0.06	1.31,-0.01,-0.01
		-0.15, 1.28,-0.03	-0.03, 1.28,-0.07	-0.08, 1.23,-0.11	-0.02, 1.23,-0.16
		0.05,-0.09, 1.15	-0.10, 0.05, 1.25	-0.11,-0.01, 1.26	-0.06,-0.06, 1.29
	< 60	1.29, 0.01,-0.08	1.27, 0.00, 0.00	1.20, 0.01,-0.02	1.24,-0.03, 0.07
		0.06, 1.19, 0.05	-0.05, 1.13,-0.05	0.07, 1.22, 0.03	0.00, 1.25, 0.00
		-0.04, 0.09, 1.22	0.02, 0.05, 1.16	0.04,-0.03, 1.24	-0.09, 0.03, 1.20

after 1 year from the recorded measurements, they had more than 50% chance of dying, and after 2 years that probability rises to over 85% with no chance of RRT.

5.3.5 Calculator

The calculator is available online here:

`https://michael-barrowman.shinyapps.io/MSCPM_for_CKD_Patients/`.

5.4 Discussion

We have used data provided by SKS to develop a Multi-State Clinical Prediction Model and then validated this model within the SKS and SERPR datasets. Within our Models, the cause of a patient's renal disease had the widest effect on patient outcomes meaning that outcomes are highly dependent on ERA-EDTA classification [228] of the diagnosis. Most groupings resulted in a lowered hazard of death and an increased hazard of RRT compared to the baseline of Systemic diseases.

The application of a Multi-state clinical prediction model to this field is novel and gives a powerful tool for providing individualised predictions of different outcomes at a wide range of time points. The model performed well under the scrutiny of validation and thus it can be considered to be reliable to predict state probabilities for patient's futures. It can also be used to estimate an expected amount of time that a patient will be on RRT. The general inclusion criteria for the development dataset, and the wide range of patient ages and measurements allows for the model to be applied to a broad spectrum of patients.

Although the inclusion criteria for SKS were broad, the demographics of the local area resulted in homogeneity of ethnicity, which may create a limitation to the applicability of our model to non-white patients, however the contribution of ethnicity to health is intertwined with deprivation level[241] and the high levels of deprivation in the catchment area of this cohort should account for this[242]. The Renal Department at SRFT is a tertiary care facility for CKD

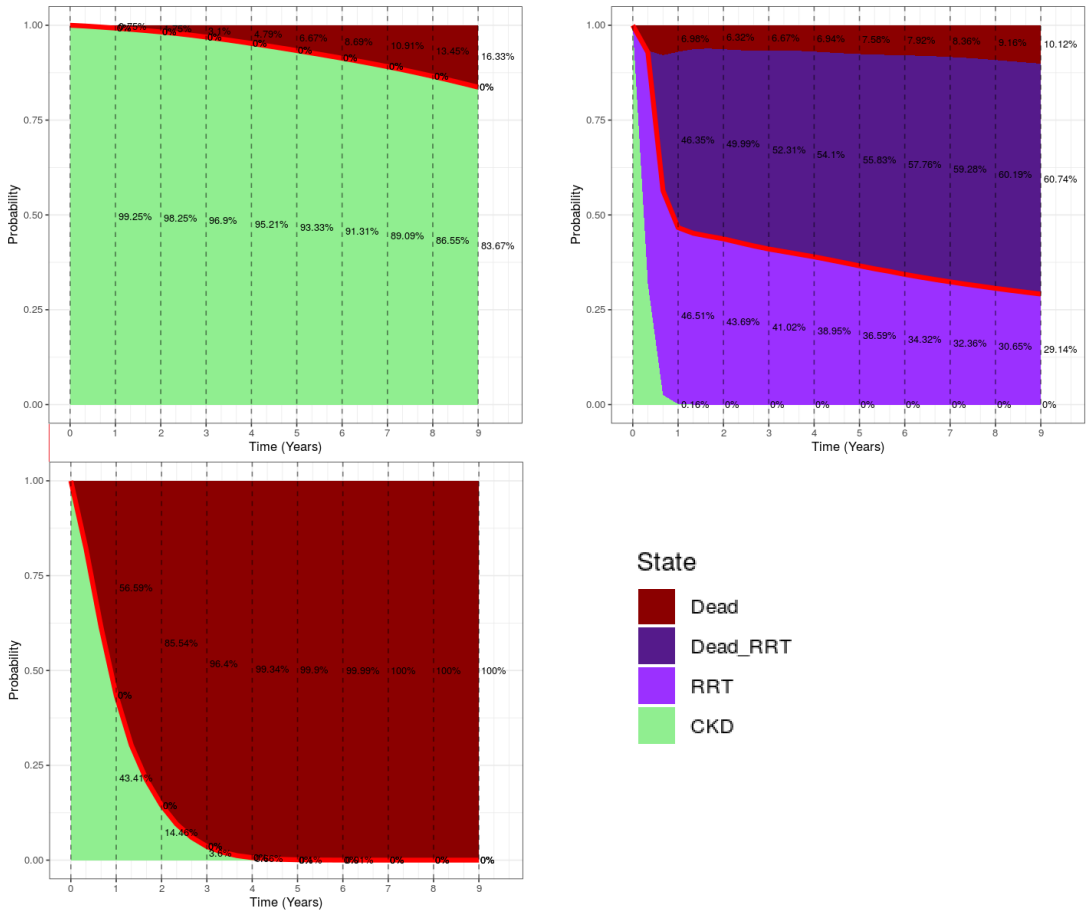

Figure 5.2: Results of Example Patients

sufferers and is well renowned for its capabilities of care meaning that it is likely to attract less-healthy patients from a wider catchment area, making the cohort of patients in the development population in worse condition than the general population of CKD patients.

There were also high levels of missingness in the eGFR and uPCR rates of changes would also produce a bias, due to these measures likely being missing not at random. The derivation of the validation dataset ensured that all patients had an eGFR Rate measurement; this was done to avoid data missing not at random (only negative or missing data would be available as patient's eGFR dropped to less than 60), however deriving data in this way could itself induce a survivor bias in the start date used for patients.

The R-P models used require the assumption that the log cumulative hazard function follow a cubic spline, however this is a very weak assumption and is assumed to be reasonable in most situations [67] and therefore we did not assess the viability of the cubic spline model. We have avoided the need for the requirements of the proportional hazards assumption by introducing time trends in the estimation of the hazard functions. This adds complexity to the model, but permits more dynamic and accurate predictions

Compared to the apparent internal validation, the model performance during the external validation was worse for all metrics. However, these results were still relatively strong which implies that the model has the potential to be transportable to a new population without much alterations being required. Due to the differences in the healthcare systems of England and Scotland, it can be appreciated that despite the populations being similar, their care would be different enough to emphasise a larger difference between our populations than that shown in our (relatively homogeneous) populations.

A key limitation of this model is that it was developed in a relatively small sample size (2,981 patients), and thus the number of patients undergoing each transition is minimal (with only 302 patients undergoing the RRT to Dead transition). Models with this weakness can often be overfit to the data they were developed in. Ordinarily, predictive models would be assessed for shrinkage and penalised appropriately if a model is overfit, however existing methods of developing shrinkage factors are not designed for MSMs and the development of a novel shrinkage factor is out of the scope of this study. It is thus reassuring that the external validation performed relatively well indicating that, with the application of suitable recalibration methods, the model could potentially be transferred to a new population and still perform well.

It is unfortunate that the performance metrics that were used to assess the calibration of the model, namely the calibration slope matrix and the calibration intercept vector (see Chapter 4 for more details), do not easily lend themselves to a graphical method. These metrics are extensions of the traditional calibration slope and intercept, respectively, which were originally graphical in nature. These original metrics were two-dimensional as they were applied to only two states (e.g. Alive vs Dead), but as the models here are applied to a four-dimensional space, it would not be feasible to visualise.

Our paper has clearly demonstrated the accuracy of such a model. It has the potential to be used as a medical tool to predict CKD patient outcomes when they are diagnosed with CKD, or at a later point on their journey. Although non-causal, predictions about patients propensity of survival can help to guide medical decisions, for example if a patient is predicted

to have a high probability of death after transitioning onto RRT, but a low probability of death without it, the patient may be better to stay off RRT. Alternatively, patients may find that they are predicted a high probability of survival if RRT is started sooner rather than later and so treatment can be guided to begin sooner. However, further research would be needed to establish the effectiveness and efficacy of this use in clinical practice [243] by comparing it to standard care and establishing whether the use of our model improves patient outcomes.

All three models produced for this work performed reasonably well in terms of accuracy, calibration and discrimination when applied internally and externally. This shows directly that the models are viable candidates for use in populations similar to both our development and our validation datasets, although further research would likely be needed to confirm their suitability such as a clinical utility assessment and perhaps further validation after recalibration.

Chapter 6

Conclusion

The key concept that this thesis aimed to address was how researchers can combine the structure of a Multi-State Model (MSM) into the field of clinical prediction modelling. Within these topics, we assessed currently available methods and provided recommendations for researchers who wish to produce a Multi-State Clinical Prediction Model (MSCPM) or to use an MSM within their research. After assessing the currently available methods, we provided novel performance metrics which can be applied to an MSCPM which are comparable to generally known metrics. By the end of the thesis, we had produced a feasible MSCPM for use in Chronic Kidney Disease (CKD) patients, as well as provided an assessment of how well this model performed under the new metrics.

The thesis can be split into 4 overarching parts

1. **The Introduction**, Chapter 1, which focuses on the current literature and describes the landscape before the start of this thesis project.

2. **The Simulation Studies**, Chapters 2 & 3, which provides a deeper analysis of some of the current techniques used within research by justifying certain techniques through the use of simulation studies.

3. **The Modelling**, Chapters 4 & 5, during which we focus on extending the current methods for model assessment and then apply these techniques to a novel model.

4. **The Conclusion**, Chapter 6, the final assessment of the thesis as a whole.

The Introduction and Conclusion are designed to act as book-ends surrounding the core of the work and the novel ideas are contained in these central four chapters. We will look at these two middle parts, The Simulation Studies and The Modelling, in turn and provide insight into their ramifications on academia in the following sections.

6.1 The Simulation Studies

Prior to this thesis, there was a gap in the literature at the intersection of competing risks analysis and causal inference. Although some authors had touched upon these topics together,

the effect that one has on the other had not been explicitly defined. When assessing the effect of a treatment (or any covariate), researchers have many options at their disposal, which largely depend on their circumstances.

When working within survival analysis, the most common method, by far, is the Cox Proportional Hazards method. This method produces an effect estimate without estimating the underlying baseline hazard and so it's simplicity is it's primary asset. However, in many circumstances, it fails to account for any competing risks that the patients may undergo. In 1999, Fine & Gray [119] wrote of a new method that accounted for these competing risks. The proportional subdistribution hazard provides effect estimates that are simple to understand in the same vein as the Cox model, but which measure the effect accounting for these competing risks.

Through simulation studies, Chapter 2 demonstrated that these confounders should always be accounted for, regardless of how strong the relationship between the confounder and either the event-of-interest or any competing events that may be present within the population. Without accounting for these biasing effects, the resulting estimates will be biased away from the true values. It also presented the work by other authors, such as Groenwold et al [164] as reasonable method of mediation of these effects.

The second simulation study of this thesis emphasises the need to account for time-dependent censoring during the assessment of calibration within clinical prediction models. The standard recommendation for this is to use the Logistic-Regression technique, however Chapter 3 demonstrated that this method is naïve to the effects that censoring has on a population.

Through the use of the Inverse-Probability-of-Censoring-Weights (IPCW), this naïve method can be augmented and improved to have a greater accuracy and provide assessments that are closer to the underlying truth. The effect that this Chapter will have on the literature is to bring the study of the calibration slope metric, in the context of the IPCW, upto similar levels as other metrics such as the Brier Score (assessed by Spitoni et al [178]) and the c-statistic (assessed by Han et al [179]); both of which agree with our results that it is far better to use the IPCW when validating models with a time-dependent aspect.

6.2 The Modelling

The following two chapters fill in the largest gap of the literature which aims to answer the core question:

How can we integrate Multi-State Models into Clinical Prediction Modelling?

As discussed in other sections of this thesis, Traditional Survival Analysis (TSA) covers the case where patients are in one of two states (usually referred to as Alive or Dead, but not always). The application of this statistical methodology to clinical prediction models is widespread and thus the design and implementation of methods to assess the quality of these models has, for the most part, become standard practice, although as discussed in Chapter 3 is far from perfect or complete.

Prior to this thesis, researchers who wished to assess the quality of a Multi-State (or Competing Risks) Clinical Prediction model would have to apply the TSA methods in creative ways.

The simplest of which was to compare only two states directly (and repeat for many combinations of states), or to compare each state against all others. These trivial extension methods are discussed in Chapter 4. However, they are simplistic and often do not give a full description of what is happening in terms of model performance.

The three most common measures that clinical prediction modellers focus on when assessing the quality of a model are the accuracy, the discrimination and the calibration of the model, and for these measures, the most common metrics are the Brier Score, the c-statistic and the calibration intercept/slope. The purpose of Chapter 4 was to extend these metrics formally and to provide researchers with a rigourous set of metrics which can easily be applied to any MSCPM.

For each of the metrics described above, the extensions were done in two ways, which take into account the two facets which may set Multi-State Modelling apart from more simple modelling. The first of these being that there are multiple states for a patient to occupy and the second being that changes can occur at different times. Other researchers have accounted for these individually, but never in combination. What's more, these extensions were done in such a way that they can easily collapse back into their original formulae if either component (additional states or time-varying) are removed.

During the process of providing these extensions, the calibration slope also provided an additional level of insight which is not present/relevant when only two states are used. This information is how well the data satisfies the Assumption of Independence of Irrelevant Alternatives. These values can help to inform whether a state can/should be removed from a model without it affecting the predictions of the other states.

Following on from the design of these performance metrics, a set of MSCPMs were constructed with an example dataset within the field of Chronic Kidney Disease (as discussed more thoroughly in Chapter 1. The models aimed to assess the patient's journey through Renal Replacement Therapy, with each model doing so at a different level of granularity.

The first was a simple two-state model, which was produced using methods similar to that of traditional survival analysis. The hazard of moving from the Alive state to the Death state was modelled using the Royston-Parmar modelling, an extension of the Cox Model, which also produces a baseline hazard using restricted cubic splines. This allows absolute predicted probabilities of transition to be produced for any arbitrary time point (within the reasonable range of the data, 10 years).

The other two models, extended from traditional survival analysis by modelling how patients move within the 'Alive' state, from being CKD sufferers to undergoing RRT and how they move onto death differently from those two states. It can be seen from the analysis in Chapter 5, that patients transitions from CKD to Death are very different from how they move from RRT to Death. However, these models are, of course, subject to bias in which patients undergo RRT and thus should not be thought of as causal effects. The final model takes the second model a step further by splitting RRT into three sub-states: Haemodialysis (HD), Peritoneal Dialysis (PD) and Kidney Transplant (Tx).

As described, these models underwent the rigourous quality assessment throughout Chapters 4 and 5 and Appendix F. These assessments included internal and external validation, which

demonstrated that the models were of reasonable quality and are thus viable candidates for progression into further study to allow them to be integrated into a healthcare setting. These models are relatively easy to apply to patients and are transparent in the way in which states relate to eachother. This gives a distinct advantage over models which use methods such as machine learning, where most of the calculations involved are unknown to the end-users.

6.3 Impact

As has been made clear throughout the individual conclusion sections in this these, the studies performed here can have large effects on other researchers work and can influence decisions made in both an academic and a clinical setting.

The work of Chapter 2 should be used by those undertaking an observational study to assess the impact of a treatment in instances where confounders may be present, particularly when that treatment is to be given to those who may be at risk of multiple causes of death. The results is that it can allow the researcher to mitigate estimates to account for the difference between the effectiveness and efficaciousness of a treatment. By following the recommendations of Groenwold [164], researchers can use these simulations to provide an empirical adjustment of their results, similar to the adjustment applied by clinical prediction model researchers when adjusting for optimism.

The work of Chapter 3 is essential when assessing calibration slope in models which have a time-dependent nature and also brings attention to the short-comings of ignoring this factor for other metrics also. Without accounting for these changes, calibration slope can drift from the underlying truth over time as the uncensored population drifts away from the original population. These drifts are ordinarily accounted for during the development of clinical prediction models, but are neglected when it comes to their validation. This means that although a model can appear to be well calibrated, over time, the accuracy of that calibration can wane.

Chapters 4 and 5 have contributed to the literature by providing essential definitions to be used when assessing the validity of multi-state models. It is the hopes of this author that having these metrics developed will demonstrate to other authors and researchers that systems such as multi-state models can be a useful and powerful technique when predicting patient outcomes. These chapters provide the formal definitions as well as justification for those definitions and their application to a real data to emphasise this. In addition the actual model developed during these chapters has been validated both internally and externally and shown to be a reasonably performing model which has the potential to be moved onto the next steps of model assessment to establish its viability in the real world.

6.4 Limitations

As with any simulation study, limitations lie in the assumptions that were made. For example, in the simulation study of Chapter 2, we made a broad assumption that the time to event data followed that of a standard hazard function, with proportional hazards; however, in reality, it is likely that the covariates would change over time. It was also a pragmatic approach to use

the same underlying hazard function for the two events, as opposed to the more complicated scenario wherein these events would have little in common.

Further, we did not model any measured confounders, apart from the treatment effect that was being estimated. For most real-world cases, there would be many covariates involved in this kind of assessment, and many of these covariates would be inter-correlated, even with any unmeasured confounding variables that are present. There is also the likelihood that some variables would have an effect on one event and not the other. We also chose to only model a single competing event, whereas, again, in reality, the populations that these theories would be applied to, could be susceptible to multiple competing events.

Similar limitations regarding our initial assumptions on the modelling scenario apply to the simulation study for Chapter 3. The main advantage of using simulation studies for this kind of work is that we know and understand the truth that is being simulated, and thus can explicitly see how and where our models differ, but in reality this would not be the case and so one would have to be careful not to 'over-correct' an already accurate model by incorrectly deriving the IPCW. It is also noteworthy that we did not include graphical methods for the assessment, which can also be used for the assessment of calibration intercept and slope using an integrated calibration index [95].

6.5 Future projects

Throughout the process of producing this thesis, there were a few avenues of study that were explored, but never came to fruition, along with several other ideas that were not feasible or within the scope of study, but that would improve the field in ways that this thesis has not quite covered. These projects have the potential to be useful and effective future pieces of research that could serve academia well, but were unable to fit into this thesis due to time/resource constraints.

One such project was a formal assessment via a scoping review (in the vein of Ng et al [244]) of the state of play of MSCPMs. A scoping review is a type of review where the reviewer does not necessarily have a specific question in mind, but seeks to assess the current literature on a specific topic [245] . Throughout the process of this thesis, many articles related to MSCPMs were sought and analysed. A scoping review would require a more formal search pattern and structure than a standard literature review (as per Chapter 1, and it would have a focus on articles which developed or validated an MSCPM, rather than those that discussed the methodology surrounding them.

As mentioned in Chapter 3, it would also be an effective piece of research to investigate the use of graphical methods which can assess model calibration. The simulations in this chapter investigated certain specific methods of assessing calibration intercept (and slope), those being the method of Kaplan-Meier, Logistic Regression and Pseudo-Observations. However, other methods, such as the Integrated Calibration Index (ICI) by Austin et al [95], could have been applied to such simulations, but would have been computationally too intensive for the scope of this chapter. Future research could be established to compare the ICI with the methods analysed herein. Further to this, we did not study the effects that particularly high levels of

censoring had on the simulated populations and this would be a useful piece of research to ensure that the weighted logistic regression method was applicable to a highly censored population.

Many of the MSCPMs found during this avenue of research were used as examples within other areas of the thesis, however these were not rigourously studied. Future research into the prevalence of such models would be beneficial to the field and could provide a useful analysis and criticism of the usage of MSCPMs. This would include describing how well guidelines (such as TRIPOD [13]) are adhered to, how many of these models have been through the formal stages of validation (whether internal and/or external) and how different models approach the methodological gaps that have been filled by this thesis. Work in this area could improve the reach of MSCPMs by making clear where they have already been used and where there are current gaps and act as a catalogue of the landscape at the time of publication. This would draw attention to fields where MSCPMs are lacking, and provide easy access and information to models that require validation.

Another avenue of study that would be beneficial for future researchers is the assessment of missing values within an MSM. For the creation of our MSCPM, we used the Multivariate Imputation for Chained Equations (MICE) as described by van Buuren & Oudshoorn [246] and its extension for survival analysis by White & Royston [75]. For this, we simply used the survival curve of all relevant transitions in order to impute missing values for any given patient. A key question during this process was whether using *all* transitional data was the most optimal method. Basic simulations were conducted, however these were not done with enough rigour in constructing realistic missingness patterns and MSM shapes to result in a finished piece of work. There will likely be a more efficient manner to re-create missing values (whether through MICE or another method specific to MSMs). Alternatively, a formal study of missingness patterns within MSMs could reveal the opposite, that the methods of White & Royston are, in fact, the best option when data is unavailable.

Ordinarily, once a model has been developed and validated, if the validation results were slightly off from perfect or the model mis-specified for a given population, it is possible to apply corrections to these models. Especially common is the corrections undergone by recalibrating a model in anticipation of its application. This is usually done by applying a linear adjustment to the predictions (on the logit scale) to bring the calibration intercept and slope to the perfect values of 1 and 0, respectively. The calibration metrics of Chapter 4 could be extended to include methods to re-calibrate a model that is mis-specified. This could be as simple as multiplying the predicted values by a matrix derived from the calibration results (a matrix which may or may not be time-dependent). This could vastly improve the currently available models and make their application much more likely in the real world. The ability to re-calibrate such models means they can potentially be applied to a much wider grouping of patients with a much more diverse demographic, and even to specilise the model to different sub-populations as needed, although these, of course, come at the cost of precision of predicting based on smaller samples of patients. Given the performance of the models developed in Chapter 5, it would be beneficial to these models if they were recalibrated using such methods to improve, say, their calibration within the <30 eGFR sub-population in the external data.

Another aspect of adjusting a model to fit a novel population is the inclusion of a shrinkage

factor. In traditional survival analysis, this would come in the form of a value which acts as a simple multiplier on the regression coefficients. This shrinkage factor helps to reduce any bias caused by overfitting of the model in the development data and sample sizes or number of predictors are often chosen to reduce the effect of a shrinkage factor. A methodology for calculating and applying a shrinkage factor to an MSM could be developed to improve utility in this field.

The development of the model in Chapter 5 used multiple packages within the R ecosystem, and there are a few packages that specialise in the use of multi-state modelling (e.g. the `msm` package [247] or the `mstate` package[248] are some of the most popular). However, as with many pieces of open source software, their utility was limited and did not cohere with the goals of this thesis. Many pieces of R code were specially written for this thesis, including the code within Appendix D, as well as the model building and derivation of the transition equations for the model and the combination of these equations into a viable predictive model. This code in combination has the potential to become a package in and of itself if generalised and documented appropriately. Again, as with much of these pieces of future work, this would bring the concept of MSCPMs to a wider audience by making it easier for researchers (particularly those without coding or statistical experience) to plug their numbers into a simple to use piece of software and produce an effective tool for clinical prediction.

The final piece of further work that would expand on this thesis would be to apply the model developed in Chapter 5 to the real-world. This would traditionally be done through some sort of impact study (as discussed in Chapter 1), wherein patients are randomly assigned to have the model applied to them or not. The knowledge of their predictions would guide treatment decisions and thereby affect their healthcare outcomes, which could then be compared to the control arm to see whether any improvement was detected. By performing an impact study of this model, it's viability in the real-world can be truly assessed to see whether it provides an overall benefit to clinicians and patients, and, thus, see whether it would be useful to proceed further in rolling it out to patients across the country. This could be performed following on from work on re-calibrating an MSCPM (as discussed above) to produce an even better predictive model.

Blank Page

Appendix A

Simulation Details

The populations used in these simulations are generated such that $\text{Corr}\,(U, Z) = \rho$ and $\text{P}\,(Z = 1) = \pi$. These baseline populations are then acted on by the event-of-interest and competing event hazard functions. These are combined into the cumulative hazard function. These are defined as follows:

$$\lambda_1(t|U, Z) = ke^{\beta_1 U + \gamma_1 Z}\lambda_0(t)\lambda_2(t|U, Z) = ke^{\beta_2 U + \gamma_2 Z}\lambda_0(t)$$

$$\Lambda(t|U, Z) = \int_0^t \lambda_1(s|U, Z) + \lambda_2(s|U, Z)\ \mathrm{d}s$$

Let ϕ and Φ be the pdf and cdf of for the standard Normal distribution respectively and I be the identity function. The process for generating the covariates and outcome data for a patient is given in Table A.1.

Table A.1: Table showing the steps taken to generate each simulated population

Step	Variable	Calculation	Description
1	x_0	$x_0 = \Phi^{-1}(\pi)$	This is the value of $x \sim N(0, 1)$ such that $\text{P}\,(x < x_0) = \pi$
2	r	$r = \frac{\rho\sqrt{\pi(1-\pi)}}{\phi(x_0)}$	This is used to ensure $\text{Corr}(U, Z) = \rho$
3	Y_1	$Y_{1i} \sim N(0, 1)$	Generate 10,000 random normal numbers
4	Y_2	$Y_{2i} \sim N(0, 1)$	Generate 10,000 random normal numbers
5	U	$U_i = Y_{1i}r + Y_{2i}\sqrt{1 - r^2}$	This defined $U \sim N(0, 1)$ such that $\text{Corr}(U, Y_1) = r$
6	Z	$Z_i = I\,(\Phi\,(Y_{1i}) < \pi)$	This defined Z such that $\text{P}\,(Z = 1) = \pi$
7	$\lambda_2\,(1)$	$\lambda_{1i}\,(t) = \lambda_1\,(t\|U_i, Z_i)$	Define an array of hazard functions
8	$\lambda_2\,(t)$	$\lambda_{2i}\,(t) = \lambda_2\,(t\|U_i, Z_i)$	Define an array of hazard functions
9	$\Lambda\,(t)$	$\Lambda_i\,(t) = \Lambda\,(t\|U_i, \|Z_i)$	Define an array of cumulative hazard functions
10	$S\,(t)$	$S_i\,(t) = \exp\,(-\Lambda_i\,(t))$	Define an array of Survival functions
11	V_1	$V_{1i} \sim Unif\,(0, 1)$	Generate 10,000 random uniform numbers
12	V_2	$V_{2i} \sim Unif\,(0, 1)$	Generate 10,000 random uniform numbers
13	T	$T_i = S_i^{-1}\,(V_{1i})$	Solve (numerically or explicitly depending on λ_0)
14	δ	$\delta_i = 1 + I\left(V_2 > \frac{\lambda_{1i}(T_i)}{\lambda_{1i}(T_i)+\lambda_{2i}(T_i)}\right)$	Generate which event occurs

due to the square root in Step 5, we restrain $|r| < 1$. This coupled with the definition of r in Step 2 means that the values of ρ and π are limited by each other such that:

$$\left| \frac{\rho\sqrt{\pi\left(1-\pi\right)}}{\phi\left(x_0\right)} \right| < 1$$

Therefore, we have bounds on ρ dependent on π:

$$|\rho| < \frac{\phi\left(x_0\right)}{\sqrt{\pi\left(1-\pi\right)}}$$

Since x_0 is dependent on π, $\phi(x_0)$ is also dependent on π and can be approximated by

$$\phi\left(x_0\right) \approx \frac{8}{5}\pi\left(1-\pi\right)$$

And thus the bound on ρ can be approximated by:

$$|\rho| < \frac{8}{5}\sqrt{\pi\left(1-\pi\right)}$$

This provides us with the bounds discussed in the main text.

Appendix B

Mathematics of Subdistribution Hazards

Due to the relationship between the cause specific hazard functions and the subdistribution hazard functions they cannot both satisfy the proportional hazards assumption. We have defined CSH functions to be proportional and so the SH functions are not. In order to find the "true" SH treatment effect for the event-of-interest, we have to find the least false parameter [158]. to do this, we must solve the following equation for $b(\Gamma_1) = 0$:

$$b\left(\Gamma_1\right) = \int_0^\infty \frac{\left(1 - F_1\left(t|Z=0\right)\right) f_1\left(t|Z=1\right) - e^{\Gamma_1}\left(1 - F_1\left(t|Z=1\right)\right) f_1\left(t|Z=0\right)}{\left(1-\pi\right)\left(1 - F_1\left(t|Z=0\right)\right) + \pi e^{\Gamma_1}\left(1 - F_1\left(t|Z=1\right)\right)} \, \mathrm{d}t$$

where f_1 and F_1 are the pdf and cdf of the subdistribution for the event-of-interest. These are defined as

$$f_1(t|Z) = \lambda_1(t|Z)S(t|Z)$$
$$F_1(t|Z) = \int_0^t f_1(u|Z)\, \mathrm{d}u$$

In populations where only a single event occurs, these are the same as the pdf and cdf of the distribution for the event. We have previously defined our hazard functions as (with $k_j = k$ if $j = 1$ and 1 otherwise)

$$\lambda_j(t|U, Z) = k_j e^{\beta_j U + \gamma_j Z} \lambda_0(t)$$

and s to find $\lambda_j(t|Z)$, we need to eliminate U from our equation by finding the expectation of U given Z.

$$\begin{aligned}
\mathrm{E}\left[e^{\beta U}|Z=0\right] &= \mathrm{E}\left[e^{\beta U}|Y_1 < x_0\right] \\
&= \mathrm{E}\left[e^{\beta}\left(rY_1 + \sqrt{1-r^2}Y_2\right)|Y_1 < x_0\right] \\
&= \mathrm{E}\left[e^{Y_2}\right]^{\beta\sqrt{1-r^2}} \times \mathrm{E}\left[e^{Y_1}|Y_1 < x_0\right]^{\beta r} \\
&= \left(e^{\frac{1}{2}}\right)^{\beta\sqrt{1-r^2}} \times \left(\int_{-\infty}^{x_0} e^{y}\left(\frac{\phi\left(y\right)}{\Phi\left(x_0\right)}\right)\,\mathrm{d}y\right)^{\beta r} \\
&= \frac{e^{\frac{\beta\sqrt{1-r^2}}{2}}}{1-\pi} \times \left(\int_{-\infty}^{x_0} e^{y}\phi\left(y\right)\,\mathrm{d}y\right)^{\beta r} \\
&= \frac{e^{\frac{\beta\sqrt{1-r^2}}{2}}}{1-\pi} \times e^{\beta r}2\left(\int_{-\infty}^{x_0} \frac{1}{\sqrt{\tau}}e^{-\frac{1}{2}(y-1)^2}\,\mathrm{d}y\right)^{\beta r} \\
&= \frac{e^{\frac{\beta}{2}\left(r+\sqrt{1-r^2}\right)}}{1-\pi} \times \left(\int_{-\infty}^{x_0-1} \frac{1}{\sqrt{\tau}}e^{-\frac{1}{2}w^2}\,\mathrm{d}w\right)^{\beta r} \\
\mathrm{E}\left[e^{\beta U}|Z=0\right] &= \frac{e^{\frac{\beta}{2}\left(r+\sqrt{1-r^2}\right)(\Phi(x_0-1))^{\beta r}}}{1-\pi}
\end{aligned}$$

Walking through each step:

1. We defined Z from Y_1 and so we can substitute $Z = 0$ with $Y_1 < x_0$ (See Appendix A).

2. We defined U based on Y_1 and Y_2 so we can substitute this formula in (See Appendix A).

3. The expectation of a product is the product of the expectations so this can be split. Similarly, powers are taken out of the expectation. Y_1 and Y_2 are independent so is dropped fro the left-hand expectations.

4. First expectation follows since $Y_2 \sim N(0, 1)$. The second is the expectation of a truncated Normal distribution.

5. By definition, $\Phi(x_0) = 1 - \pi$, so we can replace and bring it out of the integral.

6. Substituting the formula for $\phi(y)$ (Normal pdf). Note that τ is used rather than 2π to avoid notation confusion. The powers of e are combined and simplified (using completing the square) and the constant is taken out of the integral.

7. Substituting $w = y - 1$ inside the integral and not forgetting to change the limit.

8. The formula under the integration is again the Normal pdf, so we can evaluate the Normal cdf.

Similarly, for $Z = 1$, we swap the less than sign in the first line and evaluate the integral on a different range:

$$\begin{aligned}
\mathrm{E}\left[e^{\beta U}|Z=0\right] &= \mathrm{E}\left[e^{\beta U}|Y_1 < x_0\right] \\
&= \mathrm{E}\left[e^{\beta}\left(rY_1 + \sqrt{1-r^2}Y_2\right)|Y_1 < x_0\right] \\
&= \mathrm{E}\left[e^{Y_2}\right]^{\beta\sqrt{1-r^2}} \times \mathrm{E}\left[e^{Y_1}|Y_1 < x_0\right]^{\beta r} \\
&= \left(e^{\frac{1}{2}}\right)^{\beta\sqrt{1-r^2}} \times \left(\int_{x_0}^{\infty} e^{y}\left(\frac{\phi\left(y\right)}{\Phi\left(x_0\right)}\right)\,\mathrm{d}y\right)^{\beta r} \\
&= \frac{e^{\frac{\beta\sqrt{1-r^2}}{2}}}{1-\pi} \times \left(\int_{x_0}^{\infty} e^{y}\phi\left(y\right)\,\mathrm{d}y\right)^{\beta r} \\
&= \frac{e^{\frac{\beta\sqrt{1-r^2}}{2}}}{1-\pi} \times e^{\beta r}2\left(\int_{x_0}^{\infty} \frac{1}{\sqrt{\tau}}e^{-\frac{1}{2}(y-1)^2}\,\mathrm{d}y\right)^{\beta r} \\
&= \frac{e^{\frac{\beta}{2}\left(r+\sqrt{1-r^2}\right)}}{1-\pi} \times \left(\int_{x_0-1}^{\infty} \frac{1}{\sqrt{\tau}}e^{-\frac{1}{2}w^2}\,\mathrm{d}w\right)^{\beta r} \\
\mathrm{E}\left[e^{\beta U}|Z=0\right] &= \frac{e^{\frac{\beta}{2}\left(r+\sqrt{1-r^2}\right)(1-\Phi(x_0-1))^{\beta r}}}{1-\pi}
\end{aligned}$$

By defining $\pi_1 = 1 - \Phi(x_0 - 1)$, these can be combined to be:

$$\mathrm{E}\left[e^{\beta U}|Z=z\right] = \left(e^{\frac{\beta}{2}\left(r+\sqrt{1-r^2}\right)}\right)\left(\frac{\left(z\pi_1 + (1-z)\left(1-\pi_1\right)\right)^{\beta r}}{z\pi + (1-z)\left(1-\pi\right)}\right)\lambda_0(t)$$

This can then be used to produce a formula for $\Lambda(t|Z)$ and then $S(t|Z)$. From $\lambda_1(t|Z)$ and $S(t|Z)$, we can find $f_1(t|Z)$ and $F_1(t|Z)$ using the above equations and thus Γ_1 can be found. This process can be repeated to find Γ_2 by exchanging f_1 and F_1 for f_2 and F_2 and evaluation similarly.

Blank Page

Appendix C

Assessment of Calibration Slope within the IPCW analysis

C.1 Introduction

The main purpose of this paper was to assess the evaluation of calibration-in-the-large at different time points in a time-to-event clinical prediction model. Along with calibration-in-the-large, various methods of calibration can also produce measures of calibration slope. Calibration slope provides an insight into how well the model predicts outcomes across the range of predictions.

In an ideal model, the calibration slope would be 1. This implies that across the spectrum of possible predictions, the results match these prediction. If the calibration slope is below 1, then this implies that the predictions are too extreme and that the range of possible predictions would need to be tightened. If the calibration slope is greater than 1, then this implies that the predictions are too conservative and would need to be broadened to match the observations. This can be seen in Figure C.1, where the Expected values on the < 1 plot cover a broader range of values than the Observed, and so the predicted values would need to be condensed more to match the outcomes better (as well as adjusting for the calibration-in-the-large intercept value). Similarly, in the > 1 plot, the Expected values cover a narrower range of values compared to the Observed values and so predictions would need to be exaggerated to match the outcomes.

In reality, the ranges of predicted values would likely cover the most of the range from 0 to 1 and the outcome for each individual patient is binary. However the density of the predictions and outcomes would differ depending on this slope and so the inference of condensing or exaggerating the results is translatable.

C.2 Methods

The Logistic Weighted, Logistic Unweighted and Pseudo-Observation methods described in the main text can provide estimates of the calibration slope. For each of these methods, we first estimate the calibration-in-the-large as above, using a predictor as an offset, then we use this estimate as an offset to predict the calibration slope (without an intercept term). That is we

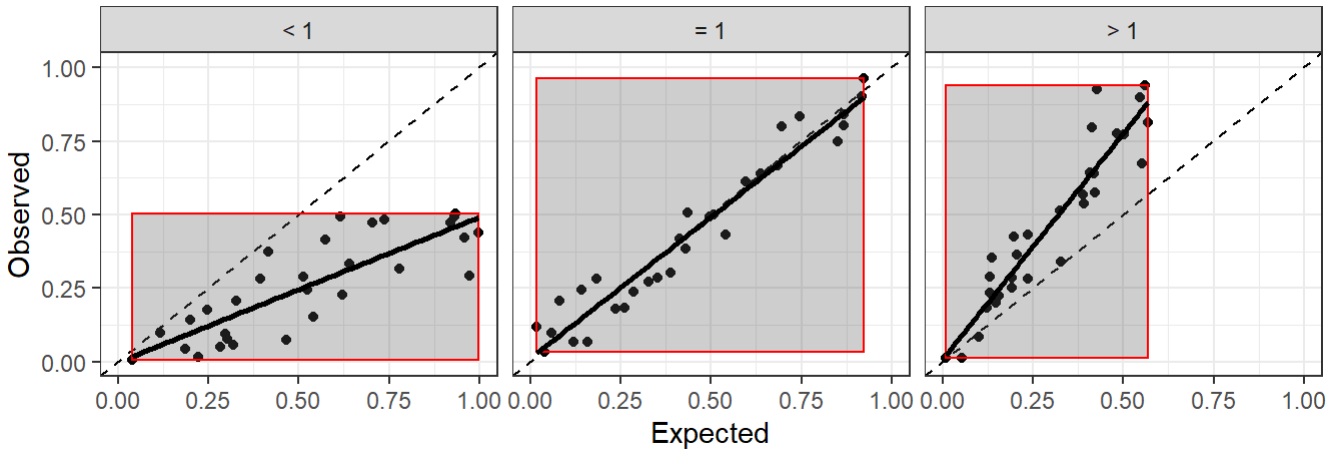

Figure C.1: Examples of low slope, perfect slope and high slope, showing the range of values

first use the appropriate methods to find $\hat{\alpha}$, where ϵ is small

$$\mathrm{O} = \hat{\alpha} + I(\mathrm{E}) + \epsilon_{\alpha}$$

Then we fix $\hat{\alpha}$ to estimate $\hat{\beta}$ below

$$\mathrm{O} = \hat{\beta}\mathrm{E} + \text{offset}\,(\hat{\alpha}) + \epsilon_{\beta}$$

As the main focus of this work is to assess the effect of these methods on the calibration-in-the-large in different models, we did not use different models than those used in the main body. Thus the slope of the three models under analysis did not change and so the expected values for these would all be 1.

C.3 Results

Results are presented similar to the main text for the same three scenarios. The full set of results is also available on the Calculator App. As discussed in the main text, the estimates are presented with time on the x-axis and the y-axis showing the performance measure, stratified by model across facets and method of analysis by colour. We will investigate the Bias, EmpSE and Coverage for the scenarios where $\beta = 1$ and $\eta = {}^{1}/_{2}$ are fixed and γ varies through -1, 0 and 1. These represent when the event and censoring are positively correlated ($\gamma = \beta = 1$), negatively correlated ($\gamma = -\beta = -1$) and when the covariate has no effect on the censoring distribution ($\gamma = 0$)

C.3.1 No Correlation

We can see from Figure C.2 that the slope for the PO measure for Bias is very large when the model does not perfectly predict the outcome and the direction of this miscalibration is dependent on whether the model Over- or Under-predicts. The LU method has a consistent

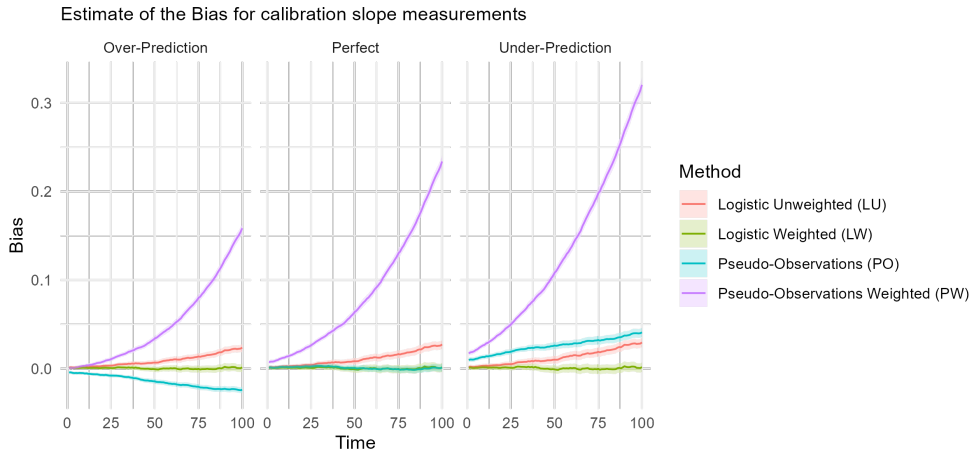

Figure C.2: Bias for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

positive bias, growing over time, regardless of which model is being analysed.

With no correlation between the covariate and the censoring, Figure C.4 shows that the EmpSE consistently increased over time, with more variability in the results for Under-Predicting model than for the Over-Predicting model. Similarly, Coverage dropped over time in all models with the Under-Predicting model dropping more significantly. The final graph in Figure C.4 shows that the Coverage for the Pseudo-Observation is lower than for the other two at the first time point and continues to steadily drop.

C.3.2 Positive Correlation

When there is a positive correlation between the covariate and the censoring of patients, we can see from Figure C.5 that the PO measurements start with a large amount of negative bias, although Figure C.6 shows relative consistency in this bias.

The LW method has the smallest amount of overall bias out of the three measures, however because all methods suffer a large amount of absolute bias, the coverage is lower than expected at almost all time point, see Figure C.7.

C.3.3 Negative Correlation

With a negative correlation, the PO method has a positive bias in the early time points of the simulations, see Figure C.8, but then reverses and has a negative bias by the end of the simulations. This causes a peak and trough effect in the Coverage, particularly strongly in the Under-Predicting Model, see Figure C.10.

The EmpSE for all methods once again increases over time, Figure C.9, and LW consistently under predicts with very low coverage throughout.

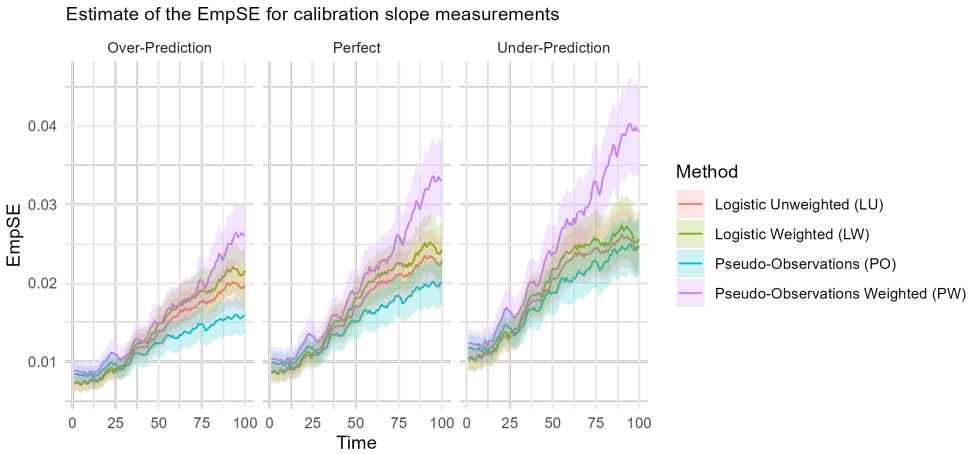

Figure C.3: EmpSE for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

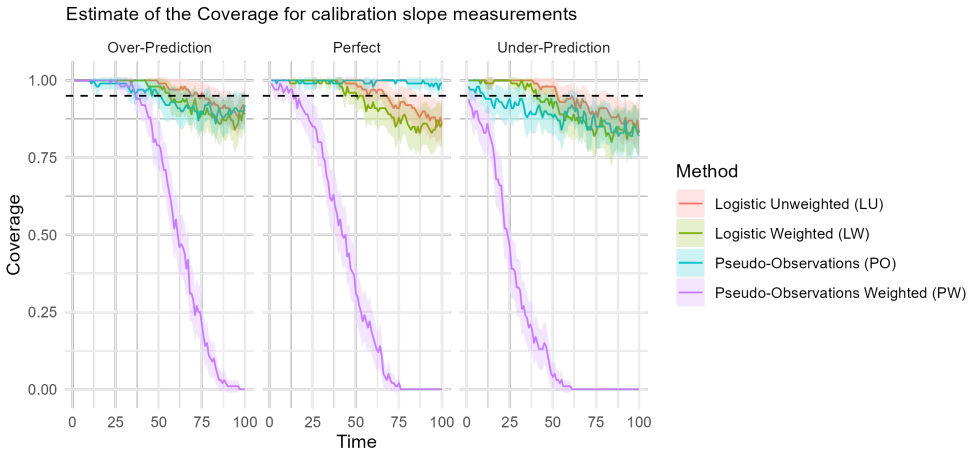

Figure C.4: Coverage for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

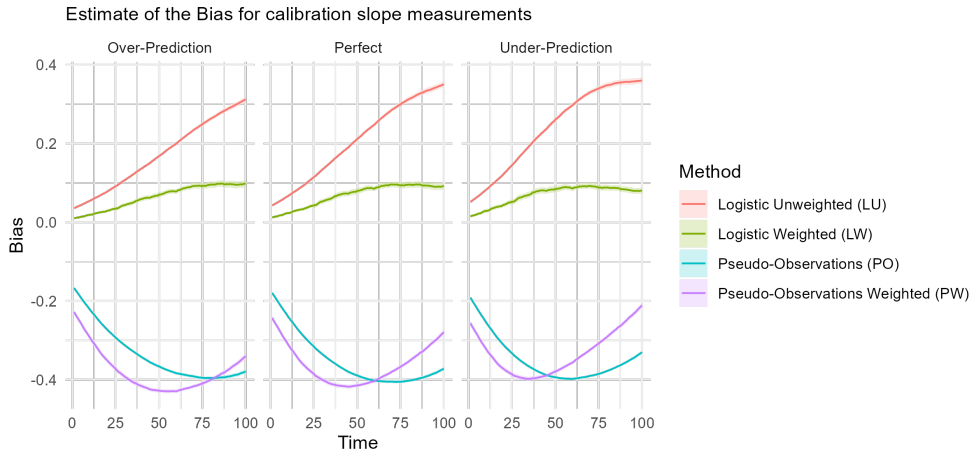

Figure C.5: Bias for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

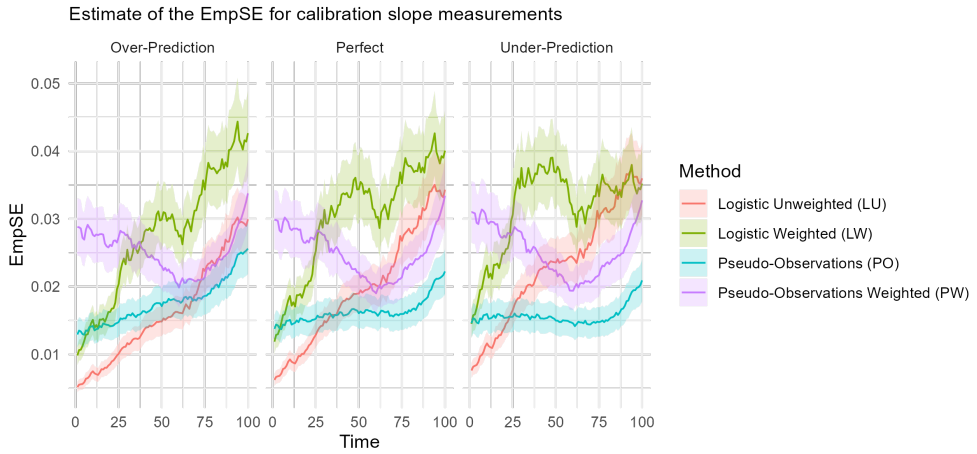

Figure C.6: EmpSE for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

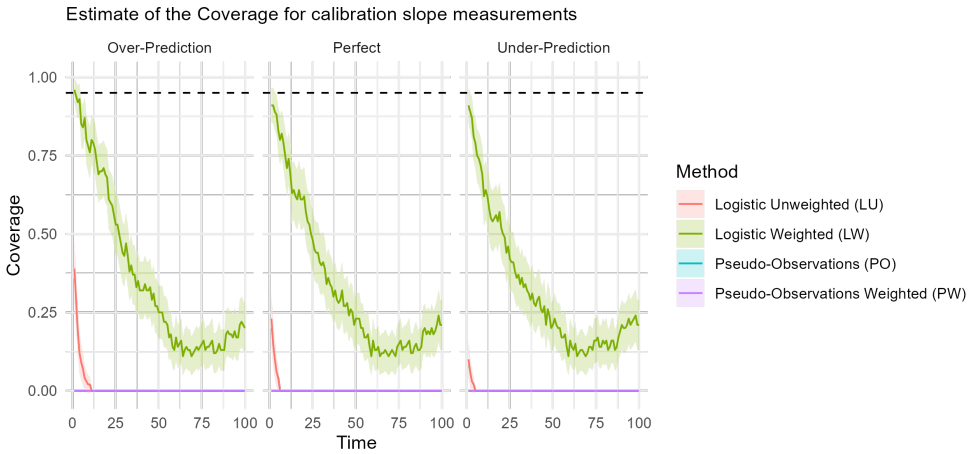

Figure C.7: Coverage for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

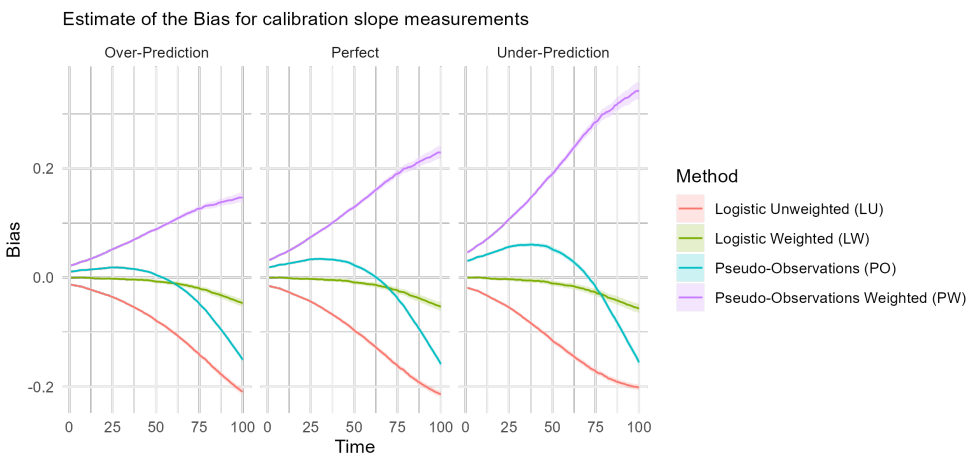

Figure C.8: Bias for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

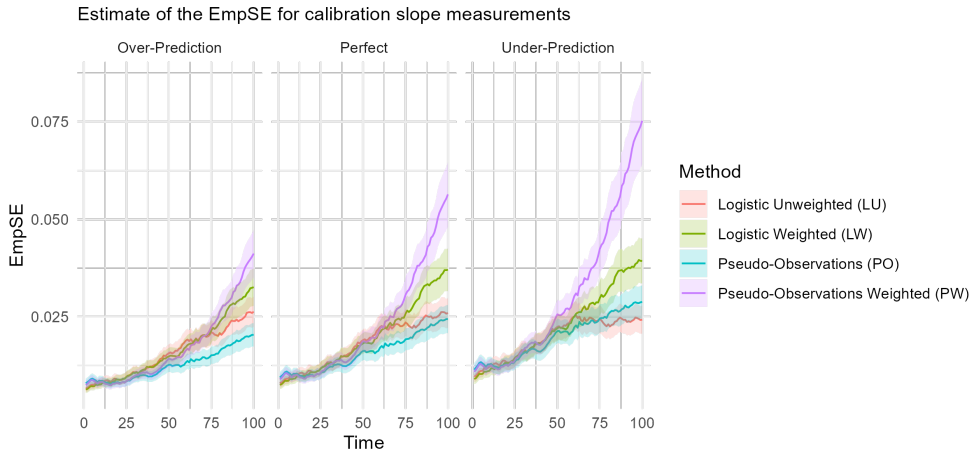

Figure C.9: EmpSE for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

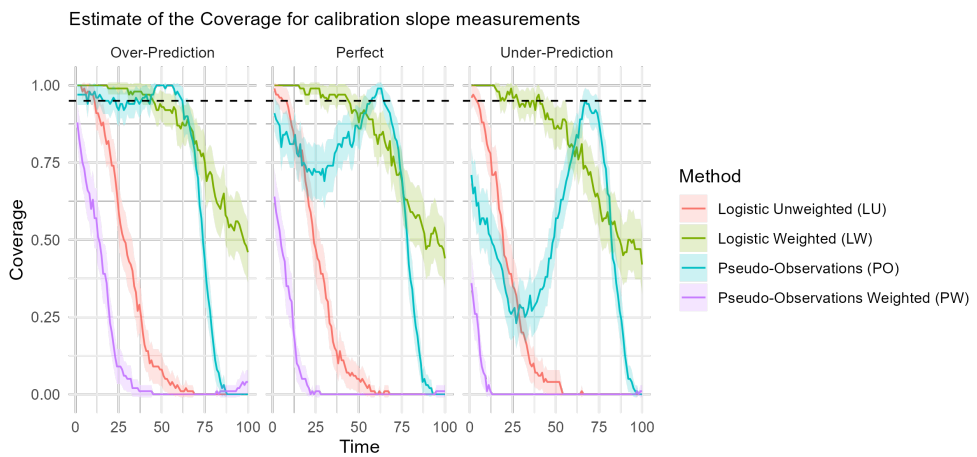

Figure C.10: Coverage for Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. 95% Confidence Intervals are included in the plot.

C.4 Discussion

As in the main text, see Chapter 3, we can see that the LW method provides the lowest overall bias compared to the other two methods of assessing Calibration. Unlike the measurements of Calibration-in-the-Large, it is still relatively inconsistent and so there will need to be further study to establish whether a better and more consistent method of measuring Calibration Slope in time-to-event studies is possible.

Appendix D

MSCPM Performance Metrics R Code

This Supplementary Material contains R functions to calculate the novel performance metrics described in the main chapter. All of these functions take five arguments:

- `states` - A character vector of states indicating the current observed state of each patient.
- `predictions` - A matrix of predictions, where each row represents a patient and each column represents a state. This matrix must have column names that match the states.
- `IPCW` - A vector of IPCW estimates for each patient at the current time point.
- `cens_name` - A character vector for which value(s) in `states` indicates a patient has been censored.
- `level` - A value to indicate what level of confidence interval should be calculated.

All three of these functions call `Check_data()`, which is defined first in this supplementary material. This function simply checks that the data (`states`, `predictions` and `IPCW`) are of the right dimensions and have the right names. The only package required to run the code is the `nnet` [202] package.

D.1 Check Function

This function checks whether the dimensions of the data are acceptable. It will throw an error if not.

```
Check_data <- function(state,predictions,IPCW,cens_name){
  N <- length(states)

  states[states %in% cens_name] <- "..censored"
```

```
  states_list <- setdiff(unique(states),"..censored")

  K <- length(states_list)


  if(N != nrow(predictions) | N != length(IPCW))
    stop("Incompatible number of patients")

  if(K != ncol(predictions))
    stop("Incompatible number of states")
}
```

D.2 Brier Score

The function to calculate the Brier Score is called `Get_Brier()`. It returns a matrix with 1 row for the BS results, and 1 row for the aBS results. The columns correspond to the actual estimate, the lower bound and the upper bound for the confidence interval.

```
Get_Brier <- function(states, predictions, IPCW, cens_name, level=0.95){

  Check_data(states,predictions,IPCW,cens_name)

  states[states %in% cens_name] <- "..censored"
  states_list <- setdiff(unique(states),"..censored")
  K <- length(states_list)

  predictions <- predictions[states %in% states_list,]
  IPCW <- IPCW[states %in% states_list]
  states <- states[states %in% states_list]

  observed <- predictions*0

  for(ik in 1:K)
    observed[,ik] <- 1*(states == colnames(observed)[ik])

  BS_i <- rowSums((predictions - observed)^2)

  BS <- weighted.mean(x = BS_i,w = IPCW)
  names(BS) <- "Est"

  BS_se <- sd(BS_i)/sqrt(length(BS_i))

  BS <- BS + c(0,-1,1)*qnorm(level)*BS_se
```

```
  names(BS) <- c("Est","lower","upper")

  Q <- vapply(X = states_list,
              FUN = function(x) weighted.mean(states == x,w=IPCW),
              FUN.VALUE = numeric(1))

  aBS <- 1-BS/sum(Q*(1-Q))
  names(aBS) <- c("Est","upper","lower")
  aBS <- aBS[c("Est","lower","upper")]

  rbind(BS,aBS)

}
```

D.3 PDI

The function to calculate the PDI is called `Get_PDI()`. It returns a matrix where the first `K` rows represent each state, and the final row is the total PDI. Each column represents the Estimate, the lower bound and the upper bound. We also follow it with a simple function to convert a PDI into a c-statistic.

```
Get_PDI <- function(states,predictions,IPCW,cens_name,level=0.95){

  Check_data(states,predictions,IPCW,cens_name)

  states[states %in% cens_name] <- "..censored"
  states_list <- setdiff(unique(states),"..censored")
  K <- length(states_list)
  N <- length(states)

  IPCW_split <- split(IPCW,states)
  IPCW_split <- IPCW_split[states_list]

  N_k <- vapply(IPCW_split,sum,numeric(1))

  V_i <- vector("list",K)
  names(V_i) <- states_list

  for(iK in 1:K){
    c_state <- state_list[[iK]]
    c_state_probs <- predictions[states == c_state,c_state]

    V_ij <- matrix(nrow=length(c_state_probs),ncol=K-1)
```

```
  for(ij in 1:(K-1)){
    jK <- setdiff(1:K,iK)[ij]
    n_state <- state_list[[jK]]
    n_state_probs <- predictions[states == n_state,c_state]
    n_state_IPCW <- IPCW_split[[n_state]]

    V_ij[,ij] <- vapply(X = c_state_probs,
                        FUN = function(x)
                          weighted.mean(x=x > n_state_probs,
                                        w=n_state_IPCW),
                        FUN.VALUE = numeric(1))
  }
  V_i[[c_state]] <- apply(V_ij,1,prod)

}

PDI_k <- mapply(weighted.mean,
                x = V_i,
                w = IPCW_split)

S_k <- mapply(
  function(V,N){
    (1/(N-1))*sum((V-PDI)^2)
  },
  V = V_i,
  N = N_k
)

se <- sqrt(S_k/N_k)
se_total <- sqrt(sum(se^2))

PDI <- mean(PDI_k)

res <- matrix(ncol=3,nrow=K+1)
rownames(res) <- c(names(PDI_k),"Total")
colnames(res) <- c("Est","lower","upper")
res[1:K,"Est"] <- PDI_k
res[1:K,"lower"] <- PDI_k - qnorm(level)*se
res[1:K,"upper"] <- PDI_k + qnorm(level)*se
res["Total","Est"] <- PDI
res["Total",c("lower","upper")] <- PDI + c(-1,1)*qnorm(level)*se_total
```

```
  res

}
```

D.4 Multinomial Calibration

The function to calculate the Multi-state extension of the Calibration Intercept and Slope is called `Get_Calib()` and takes an additional argument, `base_state` to indicate which of the states should be set as the base state for the others to be compared to. The function returns a list with two elements, `Intercept` and `Slope`. The `Intercept` component is a matrix with 1 row for each of the non-base states and a column for the Estimate, and the lower and upper bounds for the confidence interval. The `Slope` component is an array with the first two dimensions representing the non-base states and the last dimension consisting of the Estimate and the lower and upper bounds.

```
Get_Calib <- function(states,predictions,IPCW,cens_name,level=0.95,base_state){
  require(nnet)

  Check_data(states,predictions,IPCW,cens_name)

  states[states %in% cens_name] <- "..censored"
  states_list <- setdiff(unique(states),"..censored")
  K <- length(states_list)

  non_base_states <- setdiff(state_list,base_state)

  calib_data <- as.data.frame(predictions)
  calib_data$..state <- factor(states,levels=c(base_state,non_base_states))
  calib_data$..IPCW <- IPCW

  calib_data <- calib_data[states %in% states_list,]
  calib_data[,1:K] <- log(calib_data[,1:K])
  calib_data[,1:K] <- calib_data[,1:K] - calib_data[,base_state]
  calib_data[[base_state]] <- NULL


  frm <- as.formula(
    paste0("..state ~ ",
           paste0(non_base_states,
                  collapse="+")
           )
    )
```

```
  mod <- nnet::multinom(frm,data=calib_data,weights=..IPCW,trace=F)

  coefs <- coefficients(mod)

  se <- summary(mod)$standard.errors
  Intercept_se <- se[non_base_states,"(Intercept)"]

  z <- qnorm(level)

  Intercept <- matrix(nrow=K-1,ncol=3)
  colnames(Intercept) <- c("Est","lower","upper")
  rownames(Intercept) <- non_base_states
  Intercept[,"Est"] <- coefs[non_base_states,"(Intercept)"]
  Intercept[,"lower"] <- Intercept[,"Est"] - z*Intercept_se
  Intercept[,"upper"] <- Intercept[,"Est"] + z*Intercept_se

  Slope <- array(dim=c(K-1,K-1,3),
                 dimnames = list(non_base_states,
                                 non_base_states,
                                 c("Est","lower","upper")))
  Slope_se <- se[non_base_states,non_base_states]
  Slope[,,"Est"] <- coefs[non_base_states,non_base_states]
  Slope[,,"lower"] <- Slope[,,"Est"] - z*Slope_se
  Slope[,,"upper"] <- Slope[,,"Est"] + z*Slope_se


  list(Intercept = Intercept,
       Slope = Slope)

}
```

Appendix E

MSCPM Model Statistical Analysis

E.1 RP-Model

Results shown and in the main text include coefficients for both the proportional hazards (where the effect of a variable does not change over time) and time-dependent coefficients. The models were derived using a Royston-Parmer Model [67]. In it's simplest form, this model is an extension of the Cox Model, which includes the derivation of the underlying baseline hazard as a restricted cubic spline on the log-scale. However, this method of modelling also allows for variables which do not satisfy the proportional hazards assumption to be included as time-dependent covariates. For the CPM, we modelled each transition as an R-P Model, to establish a smooth hazard function from one state to another.

Firstly, we remember that the following hold, where λ is the hazard function, Λ is the cumulative hazard function and S is the probability of survival in a state at a given time:

$$\lambda(t|Z) = \frac{\mathrm{d}}{\mathrm{d}t}\Lambda(t|Z)$$
$$S(t|Z) = \exp(-\Lambda(t|Z))$$

And the R-P Model gives a smooth estimate for $S(t|X)$:

$$\log(-\log(S(t|Z)) = s(\log(t); \gamma, Z)$$

Where s is the restricted cubic spline. This restricted cubic spline is defined to have $m \geq 0$ internal knots, where $k_1, k_2, ..., k_m$ are the locations of the knots (on the log-scale). We also have two boundary knots, k_{min} and k_{max}. With $x = \log(t)$, we have:

$$s(x; \gamma, Z) = \gamma_0(Z) + \gamma_1(Z)x + \gamma_2(Z)v_1^3(x) + ... + \gamma_{m+1}(Z)v_m^3(x)$$

We further define $v_j^n(x)$ for $j > 1$:

$$v_j^n(x) = (x - k_j)_+^n - \lambda_j(x - k_{min})_+^n + (1 - \lambda_j)(x - k_{max})_+^n$$

where

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}$$

and note that, for $n \neq 0$:

$$\frac{\mathrm{d}}{\mathrm{d}x} v_j^n(x) = n v_j^{n-1}(x)$$

The γ functions are simply

$$\gamma_j(Z) = \gamma_{j0} + \gamma_{j1} Z_1 + \gamma_{j2} Z_2 + ...$$

Where each Z_i represents a patient covariate and therefore the γ_{ji} is the coefficient for that covariate associated with that piece of the cubic spline.

This combines to give the following formula for Λ:

$$\Lambda(t|Z) = \exp\left(s(\log(t); \gamma, Z)\right)$$

and therefore:

$$\lambda(t|Z) = \frac{1}{t}\Lambda(t|Z)\frac{\mathrm{d}}{\mathrm{d}x} s(\log(t); \gamma, Z)$$

where

$$\frac{\mathrm{d}}{\mathrm{d}x} s(x; \gamma, Z) = \gamma_1(Z) + 3\gamma_2(Z) v_1^2(x) + ... + 3\gamma_{m+1}(Z) v_m^2(x)$$

E.2 Multi-State Modelling

If we number the states in the models as Alive/CKD = 1, Death = 2, HD = 3, PD = 4, Tx = 5 and RRT = 6, as in Figure 5.1, then we can subscript the hazard function of the transition from state i to j as $\lambda_{ij}(t|Z)$, derived using the R-P Model in Section E.1.

We can then establish the following probabilities for our models using knowledge from Chapter 1, where $P_i(t)$ is the probability that a patient is in state i at time t, and $S_i(s, t)$ is the probability that a patient who was in state i at time s has remained in state i by time t. Since all patient start in state 1, the equivalency is $P_1(t) = S_1(0, t)$.

E.2.1 Two-State Model

$$P_1(t) = \exp\left(-\int_0^t \lambda_{12}(u)\ \mathrm{d}u\right)$$

$$P_2(t) = \int_0^t P_1(u)\lambda_{12}(u)\ \mathrm{d}u$$

E.2.2 Three-State Model

$$
\begin{aligned}
P_1(t) &= \exp\left(-\int_0^t \lambda_{12}(u) + \lambda_{16}(u)\ \mathrm{d}u\right) \\
S_6(s,t) &= \exp\left(-\int_s^t \lambda_{62}(u)\ \mathrm{d}u\right) \\
P_6(t) &= \int_0^t P_1(u)\lambda_{16}(u)S_6(u,t)\ \mathrm{d}u \\
P_2(t) &= \int_0^t P_1(u)\lambda_{12}(u)\ \mathrm{d}u + \int_0^t P_1(u)\lambda_{16}(u)\left(1 - S_6(u,t)\right)\ \mathrm{d}u
\end{aligned}
$$

E.2.3 Five-State Model

$$
\begin{aligned}
P_1(t) &= \exp\left(-\int_0^t \lambda_{12}(u) + \lambda_{13}(u) + \lambda_{14}(u) + \lambda_{15}(u)\ \mathrm{d}u\right) \\
S_3(s,t) &= \exp\left(-\int_s^t \lambda_{32}(u)\ \mathrm{d}u\right) \\
S_4(s,t) &= \exp\left(-\int_s^t \lambda_{42}(u)\ \mathrm{d}u\right) \\
P_3(t) &= \int_0^t P_1(u)\lambda_{13}(u)S_3(u,t)\ \mathrm{d}u \\
P_4(t) &= \int_0^t P_1(u)\lambda_{14}(u)S_4(u,t)\ \mathrm{d}u \\
P_5(t) &= \int_0^t P_1(u)\lambda_{15}(u)\ \mathrm{d}u \\
P_2(t) &= \int_0^t P_1(u)\lambda_{12}(u)\ \mathrm{d}u + \int_0^t P_1(u)\lambda_{13}(u)\left(1 - S_3(u,t)\right)\ \mathrm{d}u + \int_0^t P_1(u)\lambda_{14}(u)\left(1 - S_4(u,t)\right)\ \mathrm{d}u
\end{aligned}
$$

E.3 Validation Metrics

For validation purposes, we consider Death and Death after RRT/HD/PD to be distinct states meaning that for the Three-State model, we have $K = 4$ pathways a patient can take and for the Five-State model, we have $K = 7$. To compare across models, we combined states together to collapse down to simpler versions. We collapsed the Three-State model to a two-state structure by combining the CKD and RRT states into an Alive state. We collapsed the Five-State model to a three-state structure by combining the HD, PD and Tx into an RRT state and then further down to a two-state structure as with the Three-State model. We will report performance measures at 360 days (approx. 1-year), 720 days (approx. 2-years) and 1800 days (approx. 5-years). As well as presenting the performance measures over time.

The performance metrics were chosen from those defined in chapter 4.

The overall accuracy of each model was assessed using the MSM adjusted Brier Score, which is a proper score function [249] assigning 0 to a non-informative model and 1 to a perfect model,

with negative numbers implying the model performs worse than assuming every patient's state predictions are the same as the overall prevalence within the population.

The discrimination of each model was assessed using the MSM extension to the c-statistic [193]. The c-statistic is a score between 0 and 1 with higher scores suggesting a better model and a c-statistic of 0.5 suggesting the model performs no better than a non-informative model.

The calibration of each model was assessed using MSM multinomial logistic regression (MLR) [192] which extends the logistic regression to three or more mutually exclusive outcomes [8]. This produces an intercept vector of length $K-1$ and a Slope-matrix of dimension $(K-1) \times (K-1)$. As with the traditional calibration intercept for a well performing model, the MLR intercept values should all be as close to 0 as possible. The traditional calibration slope should be as close to 1 as possible and so the multi-state extension of the slope, the Slope-matrix should be as close to the identity matrix (I) as possible.

Appendix F

MSCPM Model Full Results

This Appendix shows the full results from the model described in Chapter 5. Here, we go through the results of the Two- and Five-State Models which were not discussed in Chapter 5, as well as extra results for the time-dependent covariates for the Three-State Model.

F.1 Two State Model

Table F.1 shows the proportional hazard ratios for the transitions in the Two-State Model. Older patients have a higher hazard towards death, low and decreasing eGFR increased hazard as did a history of diabetes. Patients with a primary renal diagnosis included in the ERA-EDTA [228] definition of Systemic diseases affecting the kidney had the highest likelihood of death.

Table F.1: Porportional Hazards for each transition in the Two-State Model

Var	Alive to Dead
constant term	
const	-9.072 (-20.273, 2.129)
Age	
Age	-0.009 (-0.056, 0.038)
eGFR	
eGFR	-0.066 (-0.112, -0.020)
log(eGFR Rate)	0.052 (-0.078, 0.181)
Measures	
Albumin	-0.172 (-0.322, -0.023)
Corrected Calcium	2.630 (-0.203, 5.463)
DBP	0.004 (-0.000, 0.009)
Haemoglobin	-0.011 (-0.015, -0.007)
Phosphate	1.884 (0.222, 3.546)
Gender (vs Male)	
Female	-0.168 (-0.287, -0.048)
Smoking Status (vs Former)	
Former (3y+)	-4.541 (-10.174, 1.091)
Non-Smoker	-0.648 (-1.868, 0.572)
Smoker	0.405 (-0.981, 1.791)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)	
Familial / hereditary nephropathies	0.027 (-3.640, 3.695)
Glomerular disease	0.291 (-1.476, 2.058)
Miscellaneous renal disorders	1.382 (-0.364, 3.127)
Tubulointerstitial disease	0.632 (-1.508, 2.773)
Comorbidity	
CCF	-1.471 (-2.511, -0.431)
COPD	-0.201 (-0.328, -0.075)
CVA	-0.102 (-0.265, 0.060)
DM	-0.175 (-1.205, 0.855)
IHD	0.551 (-0.572, 1.674)
LD	3.044 (-0.517, 6.606)
MI	-1.732 (-2.856, -0.607)
PVD	-1.072 (-2.134, -0.010)
ST	-1.098 (-2.256, 0.061)

The values for the γ coefficients can be seen in Table F.2. These combine together to form an individualised cubic spline for the hazard.

Table F.2: Time-Dependent γ Values for each transition in the Two-State Model

Var	Alive to Dead
Time dependent (γ_1)	
constant term	
const	1.545 (0.107, 2.982)
Age	
Age	0.008 (0.003, 0.014)
eGFR	
eGFR	0.007 (0.001, 0.012)
Measures	
Albumin	0.016 (-0.003, 0.035)
Corrected Calcium	-0.309 (-0.677, 0.059)
Phosphate	-0.199 (-0.409, 0.011)
Smoking Status (vs Former)	
Former (3y+)	0.521 (-0.185, 1.227)
Non-Smoker	0.055 (-0.099, 0.210)
Smoker	-0.005 (-0.181, 0.172)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)	
Familial / hereditary nephropathies	-0.056 (-0.516, 0.403)
Glomerular disease	-0.087 (-0.314, 0.141)
Miscellaneous renal disorders	-0.202 (-0.425, 0.020)
Tubulointerstitial disease	-0.135 (-0.400, 0.130)
Comorbidity	
CCF	0.143 (0.008, 0.277)
DM	0.042 (-0.091, 0.174)
IHD	-0.062 (-0.206, 0.082)
LD	-0.421 (-0.877, 0.036)
MI	0.202 (0.057, 0.347)
PVD	0.110 (-0.026, 0.247)
ST	0.101 (-0.048, 0.250)

Table F.3 shows the results from the internal validation in the Two-State Model. Calibration Intercept is close to 0, implying the model is well calibrated overall with a high c-statistic and Brier Score. Calibration Slope above 1 implies that the model under-estimates outcomes.

Table F.3: Internal Validation of the Two-State Model, results presented as Estimate (95% CI, where possible)

Measure	eGFR	One Year	Two Year	Five Year	Average
Brier	< 30	0.63 (0.62, 0.63)	0.69 (0.69, 0.69)	0.66 (0.66, 0.67)	0.63 (0.62, 0.63)
	< 60	0.71 (0.71, 0.72)	0.68 (0.68, 0.69)	0.66 (0.66, 0.66)	0.63 (0.63, 0.64)
c-statistic	< 30	0.82 (0.82, 0.82)	0.85 (0.84, 0.85)	0.81 (0.81, 0.81)	0.81 (0.81, 0.82)
	< 60	0.84 (0.84, 0.84)	0.83 (0.82, 0.83)	0.83 (0.82, 0.83)	0.81 (0.81, 0.81)
Intercept	< 30	0.01 (0.00, 0.01)	0.01 (0.00, 0.01)	-0.02 (-0.02, -0.01)	-0.00 (-0.01, -0.00)
	< 60	-0.02 (-0.02, -0.02)	0.00 (0.00, 0.01)	0.00 (0.00, 0.01)	-0.00 (-0.00, -0.00)
Slope	< 30	1.33	1.46	1.26	1.48
	< 60	1.23	1.25	1.30	1.51

Table F.4 shows the results from the external validation in the Two-State Model, which shows similar results to the internal validation with slightly impaired performance, which is to be expected in an external validation.

Table F.4: External Validation of the Two-State Model, results presented as Estimate (95% CI, where possible)

Measure	eGFR	One Year	Two Year	Five Year	Average
Brier	< 30	0.64 (0.63, 0.64)	0.57 (0.56, 0.57)	0.57 (0.56, 0.58)	0.56 (0.56, 0.57)
	< 60	0.67 (0.66, 0.67)	0.64 (0.63, 0.64)	0.57 (0.56, 0.57)	0.57 (0.56, 0.57)
c-statistic	< 30	0.81 (0.81, 0.82)	0.81 (0.80, 0.81)	0.80 (0.79, 0.80)	0.78 (0.78, 0.78)
	< 60	0.81 (0.81, 0.81)	0.80 (0.80, 0.81)	0.78 (0.78, 0.79)	0.78 (0.78, 0.78)
Intercept	< 30	-0.00 (-0.00, 0.00)	0.02 (0.01, 0.02)	0.00 (0.00, 0.01)	-0.00 (-0.00, 0.00)
	< 60	0.02 (0.01, 0.02)	-0.05 (-0.05, -0.04)	0.01 (0.01, 0.02)	-0.00 (-0.00, 0.00)
Slope	< 30	1.29	1.25	1.72	2.21
	< 60	1.37	1.37	2.05	1.88

F.2 Three State Model

Table F.5: Time-Dependent γ Values for each transition in the Three-State Model

Var	CKD to Dead	CKD to RRT	RRT to Dead
Time dependent (γ_1)			
constant term			
const	0.840 (-0.633, 2.314)	1.503 (1.235, 1.771)	0.975 (0.169, 1.781)
Age			
Age	0.010 (0.003, 0.016)		
eGFR			
eGFR	0.007 (0.001, 0.013)	0.058 (0.048, 0.069)	
Measures			
Albumin	0.019 (0.000, 0.037)		
Corrected Calcium	-0.276 (-0.687, 0.135)		
Phosphate	-0.186 (-0.410, 0.038)		
Gender (vs Male)			
Female	0.045 (-0.097, 0.186)		
Smoking Status (vs Former)			
Former (3y+)	0.355 (-0.371, 1.082)		218.624 (-63.849, 501.097)
Non-Smoker	0.055 (-0.104, 0.214)		0.081 (-1.117, 1.280)
Smoker	0.003 (-0.191, 0.197)		0.668 (-0.848, 2.184)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies	0.110 (-0.400, 0.619)		0.061 (-0.407, 0.528)
Glomerular disease	-0.069 (-0.289, 0.151)		-0.111 (-0.388, 0.166)
Miscellaneous renal disorders	-0.211 (-0.450, 0.028)		-0.218 (-0.589, 0.153)
Tubulointerstitial disease	-0.092 (-0.388, 0.205)		-0.242 (-0.576, 0.092)
Comorbidity			
CCF	0.180 (0.038, 0.322)		0.574 (-0.098, 1.246)

Table F.5: Time-Dependent γ Values for each transition in the Three-State Model *(continued)*

Var	CKD to Dead	CKD to RRT	RRT to Dead
MI	0.147 (0.001, 0.294)		
PVD	0.091 (-0.054, 0.237)		
ST	0.135 (-0.024, 0.294)		
Time dependent (γ_2)			
constant term			
const		0.086 (0.064, 0.108)	0.282 (0.039, 0.526)
uPCR			
uPCR		0.021 (0.007, 0.034)	
Smoking Status (vs Former)			
Former (3y+)			14.103 (-4.279, 32.486)
Non-Smoker			0.046 (-0.136, 0.228)
Smoker			0.029 (-0.168, 0.227)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies		-0.039 (-0.057, -0.020)	
Glomerular disease		0.001 (-0.012, 0.013)	
Miscellaneous renal disorders		-0.005 (-0.028, 0.017)	
Tubulointerstitial disease		-0.006 (-0.021, 0.010)	
Time dependent (γ_3)			
constant term			
const			-1.057 (-2.054, -0.061)
Time dependent (γ_4)			
constant term			
const			1.627 (-0.039, 3.292)
Time dependent (γ_5)			

Table F.5: Time-Dependent γ Values for each transition in the Three-State Model *(continued)*

Var	CKD to Dead	CKD to RRT	RRT to Dead
constant term			
const			-1.914 (-4.170, 0.343)
Smoking Status (vs Former)			
Former (3y+)			-18.897 (-47.605, 9.811)
Non-Smoker			-0.305 (-1.955, 1.344)
Smoker			-0.248 (-1.805, 1.310)
Time dependent (γ_6)			
constant term			
const			1.132 (-0.556, 2.819)
Smoking Status (vs Former)			
Former (3y+)			9.830 (-8.812, 28.471)
Non-Smoker			0.306 (-1.587, 2.198)
Smoker			0.337 (-1.432, 2.106)
Comorbidity			
CCF			0.076 (0.006, 0.146)

Table F.5 shows the γ values for the transitions from CKD to Dead, CKD to RRT and RRT to Dead. The proportional hazards and validation results for the Three-State Model can be found in Chapter 5.

F.3 Five State Model

Tables F.6, F.7, F.8 and F.9 show the proportional hazard ratios and the γ coefficients for the transitions in the Five-State Model.

Table F.6: Porportional Hazards for each transition in the Five-State Model part a

Var	CKD to Tx	HD to Dead	PD to Dead
constant term			
const	-27.179 (-41.330, -13.028)	-12.812 (-23.706, -1.918)	-36.779 (-66.159, -7.399)
Age			
Age	-0.151 (-0.224, -0.078)		0.055 (0.036, 0.073)
log(Age)	3.100 (-0.063, 6.262)	2.177 (-0.175, 4.529)	
eGFR			
eGFR	-0.157 (-0.203, -0.111)	0.011 (-0.004, 0.026)	
uPCR			
uPCR		-0.586 (-1.385, 0.214)	
Measures			
Albumin		-0.040 (-0.082, 0.002)	-0.044 (-0.094, 0.007)
Corrected Calcium			0.868 (-0.424, 2.161)
Haemoglobin			-0.022 (-0.038, -0.007)
Phosphate	0.348 (-0.420, 1.117)		
Smoking Status (vs Former)			
Former (3y+)	-5.780 (-14.562, 3.002)	-2.232 (-77.015, 72.551)	-0.465 (-4.707, 3.777)
Non-Smoker	-0.070 (-3.585, 3.445)	1.204 (-3.963, 6.371)	0.509 (-0.709, 1.727)
Smoker	-3.003 (-8.785, 2.778)	-3.057 (-8.727, 2.613)	0.567 (-0.752, 1.885)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies	-0.207 (-1.412, 0.999)	-2.191 (-9.592, 5.210)	-0.267 (-1.065, 0.532)
Glomerular disease	0.681 (-0.299, 1.661)	-0.467 (-2.134, 1.200)	-0.374 (-1.050, 0.301)
Miscellaneous renal disorders	-0.619 (-3.318, 2.080)	-0.029 (-2.096, 2.038)	1.088 (-0.106, 2.281)
Tubulointerstitial disease	-0.621 (-1.819, 0.576)	-0.613 (-3.174, 1.948)	0.555 (-0.247, 1.356)
Comorbidity			
CCF			-1.732 (-8.015, 4.550)
COPD	0.783 (0.035, 1.530)	0.283 (-0.107, 0.673)	
CVA			23.203 (-3.512, 49.918)
DM			0.587 (0.075, 1.098)
HT			-0.772 (-2.007, 0.462)
IHD			8.348 (-6.144, 22.840)
MI	1.894 (-0.095, 3.883)	-4.755 (-9.291, -0.219)	
PVD		-0.363 (-0.761, 0.035)	
ST		-0.247 (-0.645, 0.151)	

Table F.7: Porportional Hazards for each transition in the Five-State Model part b

Var	CKD to Dead	CKD to HD	CKD to PD
constant term			
const	-7.064 (-10.609, -3.518)	-6.210 (-9.773, -2.648)	-4.706 (-11.528, 2.116)

Table F.7: Porportional Hazards for each transition in the Five-State Model part b *(continued)*

Var	CKD to Dead	CKD to HD	CKD to PD
Age			
Age	-0.015 (-0.065, 0.035)	-0.035 (-0.045, -0.026)	
Age^2		-0.001 (-0.001, -0.000)	
log(Age)			-1.553 (-2.600, -0.506)
eGFR			
eGFR	-0.071 (-0.120, -0.022)	-0.626 (-0.742, -0.510)	-0.546 (-0.715, -0.377)
log(eGFR Rate)			0.505 (-0.207, 1.217)
uPCR			
log(uPCR Rate)			0.726 (-0.444, 1.897)
uPCR		1.342 (0.536, 2.149)	0.825 (-0.745, 2.395)
uPCR Rate			-0.044 (-0.139, 0.051)
Measures			
Albumin	-0.048 (-0.065, -0.031)	-0.053 (-0.085, -0.022)	-0.023 (-0.084, 0.038)
Corrected Calcium	0.360 (-0.081, 0.800)		
Haemoglobin	-0.012 (-0.016, -0.007)		
Phosphate	2.085 (0.317, 3.854)	0.517 (-0.094, 1.128)	1.064 (0.033, 2.095)
SBP		0.006 (0.001, 0.011)	0.013 (0.005, 0.022)
Gender (vs Male)			
Female	-0.225 (-0.359, -0.091)	-0.347 (-0.590, -0.104)	
Smoking Status (vs Former)			
Former (3y+)	-0.258 (-0.922, 0.407)	-0.368 (-1.230, 0.493)	-0.386 (-1.540, 0.768)
Non-Smoker	-0.180 (-0.326, -0.034)	-0.190 (-0.450, 0.070)	-0.038 (-0.383, 0.307)
Smoker	0.354 (0.160, 0.548)	0.257 (-0.060, 0.573)	0.427 (0.046, 0.807)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies	-0.404 (-0.825, 0.017)	-3.618 (-6.932, -0.304)	0.087 (-8.414, 8.589)
Glomerular disease	-0.352 (-0.584, -0.119)	-0.760 (-2.889, 1.370)	2.126 (-3.392, 7.643)
Miscellaneous renal disorders	-0.229 (-0.476, 0.017)	-0.209 (-3.630, 3.213)	6.568 (-0.361, 13.496)
Tubulointerstitial disease	-0.424 (-0.697, -0.151)	-1.157 (-3.716, 1.401)	5.554 (-0.575, 11.682)
Comorbidity			
CCF	-0.364 (-0.503, -0.226)		
COPD	-0.284 (-0.426, -0.143)		
DM	0.120 (-0.011, 0.251)		
MI	-1.714 (-2.780, -0.648)		
PVD	-0.234 (-0.377, -0.092)		
ST	-0.290 (-0.446, -0.134)	-0.445 (-0.770, -0.120)	

Table F.8: Time-Dependent γ Values for each transition in the Five-State Model part a

Var	CKD to Tx	HD to Dead	PD to Dead
Time dependent (γ_1)			
constant term			
const	1.547 (0.540, 2.553)	0.727 (-0.002, 1.457)	8.631 (1.998, 15.265)
Smoking Status (vs Former)			
Former (3y+)	0.864 (-0.293, 2.020)	-0.183 (-13.681, 13.315)	
Non-Smoker	-0.009 (-0.472, 0.454)	-0.277 (-1.256, 0.702)	
Smoker	0.319 (-0.441, 1.080)	0.753 (-0.320, 1.826)	
Comorbidity			
CCF			0.366 (-0.888, 1.620)
CVA			-5.462 (-11.477, 0.554)
IHD			-2.100 (-5.441, 1.241)
MI		0.984 (0.087, 1.881)	
Time dependent (γ_2)			
constant term			
const	-0.674 (-1.237, -0.111)	0.395 (0.053, 0.737)	1.162 (0.352, 1.971)
Comorbidity			
CVA			-0.355 (-0.907, 0.196)
IHD			-0.493 (-0.932, -0.055)
Time dependent (γ_3)			
constant term			
const	1.081 (0.344, 1.818)	-1.574 (-3.117, -0.032)	-1.611 (-3.344, 0.121)
eGFR			
eGFR	-0.004 (-0.007, -0.001)		
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies	-0.144 (-0.252, -0.035)	1.758 (-5.355, 8.871)	
Glomerular disease	0.037 (-0.042, 0.116)	0.267 (-1.389, 1.923)	
Miscellaneous renal disorders	0.064 (-0.153, 0.282)	0.763 (-1.368, 2.894)	
Tubulointerstitial disease	-0.030 (-0.111, 0.050)	-5.412 (-15.642, 4.819)	
Time dependent (γ_4)			
constant term			
const		3.087 (0.179, 5.994)	2.644 (-0.790, 6.077)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			

Table F.8: Time-Dependent γ Values for each transition in the Five-State Model part a*(continued)*

Var	CKD to Tx	HD to Dead	PD to Dead
Familial / hereditary nephropathies		-6.645 (-24.740, 11.451)	
Glomerular disease		-2.286 (-7.150, 2.579)	
Miscellaneous renal disorders		-4.262 (-11.574, 3.051)	
Tubulointerstitial disease		18.529 (-15.551, 52.609)	
Comorbidity			
MI		0.054 (-0.005, 0.112)	
Time dependent (γ_5)			
constant term			
const		-3.577 (-7.134, -0.021)	-7.683 (-16.560, 1.193)
Smoking Status (vs Former)			
Former (3y+)			7.548 (-0.830, 15.927)
Non-Smoker			0.755 (-0.857, 2.366)
Smoker			-0.085 (-1.699, 1.530)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies		8.753 (-10.491, 27.998)	
Glomerular disease		5.130 (-2.078, 12.338)	
Miscellaneous renal disorders		8.176 (-4.221, 20.574)	
Tubulointerstitial disease		-26.816 (-73.441, 19.808)	
Comorbidity			
CCF			-1.796 (-5.061, 1.469)
IHD			6.534 (-0.283, 13.352)
Time dependent (γ_6)			
constant term			
const		2.102 (-0.215, 4.420)	6.833 (-3.221, 16.886)
Age			
log(Age)		-0.079 (-0.177, 0.020)	
Smoking Status (vs Former)			
Former (3y+)		-0.147 (-1.371, 1.077)	-10.486 (-22.059, 1.087)
Non-Smoker		-0.026 (-0.133, 0.082)	-1.005 (-3.209, 1.199)
Smoker		0.091 (-0.025, 0.207)	0.140 (-2.067, 2.346)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies		-4.062 (-13.538, 5.413)	

Table F.8: Time-Dependent γ Values for each transition in the Five-State Model part a*(continued)*

Var	CKD to Tx	HD to Dead	PD to Dead
Glomerular disease		-3.652 (-8.701, 1.398)	
Miscellaneous renal disorders		-5.414 (-14.068, 3.239)	
Tubulointerstitial disease		15.910 (-10.320, 42.139)	
Comorbidity			
CCF			2.565 (-1.886, 7.017)
CVA			0.314 (-0.629, 1.257)
IHD			-7.802 (-16.654, 1.051)

Table F.9: Time-Dependent γ Values for each transition in the Five-State Model part b

Var	CKD to Dead	CKD to HD	CKD to PD
Time dependent (γ_1)			
constant term			
const	1.262 (0.839, 1.686)	1.450 (0.892, 2.009)	1.597 (1.006, 2.188)
Age			
Age	0.009 (0.003, 0.016)		
eGFR			
eGFR	0.008 (0.001, 0.014)	0.068 (0.054, 0.082)	0.056 (0.035, 0.077)
Measures			
Phosphate	-0.216 (-0.440, 0.009)		
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies		0.619 (0.165, 1.073)	0.052 (-1.524, 1.628)
Glomerular disease		0.066 (-0.213, 0.344)	-0.419 (-1.433, 0.595)
Miscellaneous renal disorders		-0.023 (-0.468, 0.421)	-1.499 (-2.686, -0.312)
Tubulointerstitial disease		0.097 (-0.232, 0.427)	-1.129 (-2.238, -0.019)
Comorbidity			
MI	0.196 (0.057, 0.335)		
Time dependent (γ_2)			
constant term			
const		-0.041 (-0.191, 0.109)	0.055 (0.001, 0.109)
Measures			
Phosphate			0.026 (-0.002, 0.054)
Primary Renal Diagnosis (vs Systemic diseases affecting the kidney)			
Familial / hereditary nephropathies			-0.024 (-0.137, 0.088)
Glomerular disease			-0.025 (-0.092, 0.042)
Miscellaneous renal disorders			-0.101 (-0.191, -0.011)
Tubulointerstitial disease			-0.074 (-0.148, -0.000)

Table F.9: Time-Dependent γ Values for each transition in the Five-State Model part b*(continued)*

Var	CKD to Dead	CKD to HD	CKD to PD
Time dependent (γ_3)			
constant term			
const		0.179 (-0.004, 0.362)	
uPCR			
uPCR		0.030 (0.002, 0.058)	

Tables F.10 and F.12 shows the results from the internal validation in the Five-State Model. Table F.10 and F.12 shows the results from the external validation in the Five-State Model.

Table F.10: External Validation of the Five-State Model, results presented as Estimate (95% CI)

eGFR	Measure	One Year	Two Year	Five Year	Average
< 30	Brier	0.70 (0.70, 0.71)	0.72 (0.71, 0.72)	0.64 (0.64, 0.65)	0.63 (0.63, 0.64)
< 60	Brier	0.73 (0.72, 0.73)	0.71 (0.70, 0.71)	0.67 (0.67, 0.68)	0.64 (0.64, 0.65)
< 30	c-statistic	0.85 (0.85, 0.85)	0.85 (0.84, 0.85)	0.82 (0.82, 0.82)	0.82 (0.81, 0.82)
< 60	c-statistic	0.85 (0.85, 0.85)	0.83 (0.82, 0.83)	0.82 (0.82, 0.82)	0.81 (0.81, 0.82)
< 30	Intercept	-0.01 (-0.01, -0.01)	0.01 (0.00, 0.01)	0.00 (0.00, 0.01)	-0.00 (-0.00, 0.00)
		-0.02 (-0.02, -0.01)	-0.01 (-0.02, -0.01)	-0.00 (-0.01, -0.00)	-0.00 (-0.01, -0.00)
		-0.02 (-0.02, -0.01)	-0.01 (-0.02, -0.01)	-0.02 (-0.03, -0.02)	0.00 (-0.00, 0.00)
		0.04 (0.03, 0.04)	-0.02 (-0.03, -0.02)	0.01 (0.01, 0.02)	-0.00 (-0.00, 0.00)
		0.00 (0.00, 0.00)	-0.01 (-0.01, -0.00)	0.00 (-0.00, 0.00)	-0.00 (-0.01, -0.00)
		-0.01 (-0.01, -0.00)	0.01 (0.00, 0.01)	0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
< 60	Intercept	0.02 (0.01, 0.02)	0.02 (0.02, 0.03)	0.03 (0.02, 0.03)	0.00 (-0.00, 0.00)
		-0.01 (-0.01, -0.00)	-0.02 (-0.02, -0.01)	0.00 (0.00, 0.01)	-0.00 (-0.00, -0.00)
		0.02 (0.01, 0.02)	-0.02 (-0.02, -0.01)	-0.02 (-0.02, -0.01)	0.00 (-0.00, 0.00)
		-0.02 (-0.03, -0.02)	-0.00 (-0.00, -0.00)	-0.00 (-0.01, -0.00)	0.00 (-0.00, 0.00)
		0.03 (0.02, 0.03)	0.02 (0.02, 0.02)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, -0.00)
		0.01 (0.01, 0.01)	-0.01 (-0.01, -0.00)	0.02 (0.01, 0.02)	-0.00 (-0.00, 0.00)

Table F.11: Internal Validation of the Five-State Model, results presented as Estimate (95% CI)

eGFR	Measure	One Year	Two Year	Five Year	Average
< 30	Brier	0.74 (0.74, 0.75)	0.72 (0.72, 0.72)	0.67 (0.66, 0.67)	0.69 (0.69, 0.69)
< 60	Brier	0.76 (0.75, 0.76)	0.71 (0.71, 0.72)	0.65 (0.65, 0.66)	0.68 (0.68, 0.69)
< 30	c-statistic	0.88 (0.88, 0.88)	0.86 (0.85, 0.86)	0.83 (0.83, 0.84)	0.84 (0.84, 0.84)
< 60	c-statistic	0.88 (0.87, 0.88)	0.87 (0.87, 0.87)	0.86 (0.86, 0.86)	0.84 (0.84, 0.85)
< 30	Intercept	0.00 (-0.00, 0.00)	0.01 (0.00, 0.01)	-0.01 (-0.01, -0.00)	0.00 (-0.00, 0.00)
		-0.01 (-0.01, -0.01)	0.01 (0.01, 0.01)	-0.00 (-0.01, -0.00)	0.00 (-0.00, 0.00)
		-0.01 (-0.01, -0.01)	-0.00 (-0.00, 0.00)	-0.01 (-0.01, -0.00)	-0.00 (-0.00, 0.00)
		0.01 (0.00, 0.01)	-0.00 (-0.00, -0.00)	-0.04 (-0.05, -0.04)	-0.00 (-0.00, 0.00)
		-0.00 (-0.00, 0.00)	0.00 (0.00, 0.01)	0.01 (0.00, 0.01)	-0.00 (-0.00, 0.00)
		-0.01 (-0.01, -0.00)	0.00 (0.00, 0.00)	-0.00 (-0.00, -0.00)	0.00 (0.00, 0.01)
< 60	Intercept	-0.02 (-0.02, -0.02)	0.00 (0.00, 0.00)	-0.00 (-0.00, 0.00)	-0.00 (-0.01, -0.00)
		0.00 (0.00, 0.01)	0.00 (0.00, 0.00)	0.00 (0.00, 0.01)	0.00 (0.00, 0.00)
		-0.00 (-0.01, -0.00)	-0.01 (-0.02, -0.01)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
		-0.00 (-0.01, -0.00)	-0.02 (-0.02, -0.01)	-0.01 (-0.01, -0.01)	-0.00 (-0.00, 0.00)
		0.01 (0.01, 0.01)	-0.00 (-0.00, -0.00)	-0.01 (-0.01, -0.01)	0.00 (-0.00, 0.00)
		0.00 (0.00, 0.00)	-0.00 (-0.00, -0.00)	-0.00 (-0.00, -0.00)	-0.00 (-0.00, 0.00)

Table F.12: External Validation of the Five-State Model, Calibration Slope results presented as Estimate only

eGFR	One Year	Two Year
< 30	1.16, 0.02, -0.00, -0.01, 0.00, 0.06	1.25, 0.02, 0.01, 0.02, -0.07, 0.02
	-0.03, 1.08, -0.00, -0.00, 0.04, -0.04	0.00, 1.39, 0.08, -0.03, 0.02, 0.05
	-0.05, -0.02, 1.18, -0.04, -0.05, 0.02	-0.01, -0.01, 1.26, 0.04, -0.04, 0.03
	-0.01, -0.04, -0.06, 1.17, 0.00, -0.00	-0.00, -0.00, -0.10, 1.21, -0.04, 0.02
	-0.01, -0.02, 0.00, -0.01, 1.14, -0.04	-0.02, -0.02, 0.02, -0.08, 1.20, -0.00
	0.03, -0.00, -0.03, 0.00, 0.01, 1.23	-0.04, 0.00, -0.05, 0.02, -0.06, 1.44
< 60	1.31, 0.02, 0.05, -0.00, -0.01, -0.00	1.39, -0.04, 0.00, -0.00, 0.01, -0.01
	0.02, 1.16, 0.04, 0.01, 0.04, -0.03	-0.06, 1.31, 0.03, 0.00, 0.02, -0.06
	0.04, 0.01, 1.08, 0.01, 0.01, -0.04	0.00, -0.02, 1.14, 0.01, -0.04, 0.04
	-0.04, -0.04, -0.01, 1.08, -0.00, -0.00	0.01, -0.02, 0.00, 1.37, -0.03, 0.00
	-0.02, 0.04, -0.03, -0.03, 1.27, -0.00	0.04, -0.02, -0.00, -0.01, 1.34, 0.02
	-0.01, 0.02, 0.03, -0.02, 0.07, 1.16	-0.02, 0.02, -0.03, -0.02, -0.01, 1.32
	Five Year	Average
< 30	1.45, -0.04, 0.06, -0.07, 0.00, 0.04	1.46, 0.01, -0.00, 0.00, -0.01, 0.00
	-0.01, 1.31, 0.04, 0.01, -0.01, 0.03	0.01, 1.46, 0.01, 0.00, -0.00, -0.00
	0.01, 0.06, 1.50, 0.01, 0.02, -0.00	-0.00, 0.00, 1.54, -0.01, 0.00, -0.00
	-0.01, 0.00, 0.08, 1.56, 0.02, 0.01	0.00, -0.02, 0.00, 1.51, -0.00, 0.01
	-0.09, 0.01, -0.02, -0.06, 1.47, -0.00	-0.01, 0.00, 0.00, -0.00, 1.47, -0.00
	-0.03, 0.04, 0.07, 0.08, 0.01, 1.35	0.00, -0.00, -0.00, -0.00, -0.01, 1.47
< 60	1.32, 0.04, -0.04, -0.01, -0.02, 0.03	1.49, -0.00, -0.00, -0.00, -0.01, 0.01
	-0.06, 1.21, 0.01, 0.00, -0.06, -0.01	-0.01, 1.45, -0.00, 0.00, 0.00, -0.00
	0.01, -0.08, 1.70, 0.04, -0.03, -0.05	0.00, -0.00, 1.44, 0.01, -0.00, -0.01
	-0.01, 0.03, -0.02, 1.35, -0.02, 0.02	0.00, -0.00, 0.01, 1.47, -0.00, 0.00
	-0.05, -0.03, 0.03, -0.04, 1.43, 0.07	-0.00, -0.00, -0.00, 0.00, 1.47, 0.01
	0.01, 0.04, -0.04, -0.01, -0.00, 1.28	0.00, -0.00, -0.00, -0.00, 0.00, 1.44

Table F.13: Internal Validation of the Five-State Model, Calibration Slope results presented as Estimate only

eGFR	One Year	Two Year
< 30	1.14, -0.01, 0.00, 0.02, 0.00, 0.01	1.42, -0.05, 0.03, -0.01, -0.05, 0.00
	-0.00, 1.09, -0.05, -0.01, 0.01, 0.00	0.02, 1.12, 0.03, -0.01, 0.03, -0.00
	-0.04, 0.06, 1.22, 0.02, 0.03, 0.00	0.01, -0.01, 1.22, -0.00, 0.00, 0.04
	-0.06, 0.02, 0.01, 1.25, 0.00, -0.02	-0.03, 0.00, -0.05, 1.21, -0.01, 0.00
	-0.02, 0.03, 0.02, -0.02, 1.13, 0.00	-0.02, -0.03, -0.00, -0.01, 1.39, -0.00
	0.05, -0.04, -0.06, -0.02, -0.00, 1.30	-0.03, 0.04, 0.04, -0.00, 0.03, 1.15
< 60	1.07, 0.03, 0.02, 0.02, 0.01, -0.00	1.14, -0.02, 0.04, 0.02, 0.00, -0.09
	0.00, 1.18, 0.01, 0.00, -0.06, 0.01	-0.01, 1.10, 0.00, -0.07, -0.01, -0.00
	0.01, 0.00, 1.03, 0.00, -0.05, 0.00	-0.03, 0.02, 1.12, 0.03, 0.00, -0.04
	-0.04, -0.03, -0.03, 1.11, 0.01, -0.03	0.00, 0.02, -0.03, 1.11, -0.02, 0.01
	0.00, -0.03, -0.02, -0.02, 1.15, 0.01	-0.00, -0.06, -0.05, 0.02, 1.11, -0.03
	0.02, -0.01, 0.00, -0.00, -0.01, 1.11	0.07, -0.01, -0.03, 0.02, 0.04, 1.18
	Five Year	Average
< 30	1.22, -0.00, 0.05, -0.00, -0.05, 0.05	1.28, 0.00, -0.00, 0.01, 0.00, 0.00
	-0.04, 1.11, -0.01, 0.03, 0.04, 0.03	0.00, 1.28, 0.00, 0.00, -0.00, 0.01
	0.02, -0.03, 1.24, -0.03, -0.03, 0.01	0.00, 0.01, 1.25, 0.00, -0.00, 0.00
	0.02, -0.03, -0.00, 1.20, -0.05, 0.01	-0.00, 0.00, 0.00, 1.28, -0.00, 0.00
	0.00, 0.08, -0.00, 0.01, 1.25, -0.06	-0.00, 0.00, -0.00, 0.01, 1.26, -0.00
	0.01, -0.00, -0.00, 0.03, -0.05, 1.14	-0.00, 0.01, 0.00, -0.00, -0.00, 1.31
< 60	1.27, 0.02, 0.02, -0.03, 0.04, -0.06	1.31, 0.01, -0.00, 0.00, 0.00, 0.00
	-0.00, 1.20, -0.00, 0.00, 0.04, -0.00	-0.00, 1.29, -0.00, 0.00, -0.00, 0.00
	-0.03, -0.02, 1.22, 0.02, 0.05, 0.03	-0.00, 0.00, 1.30, -0.00, -0.00, -0.00
	0.00, -0.02, -0.01, 1.30, -0.00, 0.04	0.00, -0.00, -0.00, 1.28, 0.00, -0.01
	0.00, -0.05, 0.00, -0.05, 1.31, 0.01	0.00, 0.00, 0.00, 0.00, 1.27, -0.01
	-0.01, -0.04, -0.05, 0.02, -0.00, 1.19	0.01, -0.00, -0.00, -0.00, 0.00, 1.28

Blank Page

References

[1] Harry Hemingway, Peter Croft, Pablo Perel, Jill A. Hayden, Keith Abrams, Adam Timmis, Andrew Briggs, Ruzan Udumyan, Karel G. M. Moons, Ewout W. Steyerberg, Ian Roberts, Sara Schroter, Douglas G. Altman, and Richard D. Riley. Prognosis Research Strategy (PROGRESS) 1: A Framework for Researching Clinical Outcomes. *BMJ*, 346:e5595, February 2013.

[2] Aroon D. Hingorani, Daniëlle A. van der Windt, Richard D. Riley, Keith Abrams, Karel G. M. Moons, Ewout W. Steyerberg, Sara Schroter, Willi Sauerbrei, Douglas G. Altman, and Harry Hemingway. Prognosis Research Strategy (PROGRESS) 4: Stratified Medicine Research. *BMJ*, 346:e5793, February 2013.

[3] Ewout W. Steyerberg, Karel G. M. Moons, Danielle A. van der Windt, Jill A. Hayden, Pablo Perel, Sara Schroter, Richard D. Riley, Harry Hemingway, Douglas G. Altman, and for the PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLOS Medicine*, 10(2):e1001381, February 2013.

[4] Hippocrates and Francis Adams. *The Genuine Works of Hippocrates*. New York, W. Wood and company, 1886.

[5] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. Development and Validation of QRISK3 Risk Prediction Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study. *BMJ*, 357, May 2017.

[6] Kristian Thygesen, Joseph S. Alpert, Harvey D. White, and Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction. Universal Definition of Myocardial Infarction. *Journal of the American College of Cardiology*, 50(22):2173–2195, November 2007.

[7] Probst, Veltmann C., Eckardt L., Meregalli P.G., Gaita F., Tan H.L., Babuty D., Sacher F., Giustetto C., Schulze-Bahr E., Borggrefe M., Haissaguerre M., Mabo P., Le Marec H., Wolpert C., and Wilde A.A.M. Long-Term Prognosis of Patients Diagnosed With Brugada Syndrome. *Circulation*, 121(5):635–643, February 2010.

[8] Richard D. Riley, Danielle van der Windt, Peter Croft, and Karel G. M. Moons. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford University Press, first edition, 2019.

[9] Richard D. Riley, Jill A. Hayden, Ewout W. Steyerberg, Karel G. M. Moons, Keith Abrams, Panayiotis A. Kyzas, Núria Malats, Andrew Briggs, Sara Schroter, Douglas G. Altman, and Harry Hemingway. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Medicine*, 10(2), February 2013.

[10] Patrick Royston, Karel G. M. Moons, Douglas G. Altman, and Yvonne Vergouwe. Prognosis and Prognostic Research: Developing a Prognostic Model. *BMJ*, 338:b604, March 2009.

[11] Douglas G. Altman, Yvonne Vergouwe, Patrick Royston, and Karel G. M. Moons. Prognosis and Prognostic Research: Validating a Prognostic Model. *BMJ*, 338:b605, May 2009.

[12] Karel G. M. Moons, Douglas G. Altman, Yvonne Vergouwe, and Patrick Royston. Prognosis and Prognostic Research: Application and Impact of Prognostic Models in Clinical Practice. *BMJ*, 338:b606, June 2009.

[13] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel GM Moons. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, 13(1):1, January 2015.

[14] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, and Gary S. Collins. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, 162(1):W1, January 2015.

[15] Richard D. Riley, Joie Ensor, Kym I. E. Snell, Thomas P. A. Debray, Doug G. Altman, Karel G. M. Moons, and Gary S. Collins. External Validation of Clinical Prediction Models Using Big Datasets from E-Health Records or IPD Meta-Analysis: Opportunities and Challenges. *BMJ*, 353, June 2016.

[16] Gary S. Collins, Omar Omar, Milensu Shanyinde, and Ly-Mee Yu. A Systematic Review Finds Prediction Models for Chronic Kidney Disease Were Poorly Reported and Often Developed Using Inappropriate Methods. *Journal of Clinical Epidemiology*, 66(3):268–277, March 2013.

[17] E.W. Steyerberg. Overfitting and Optimism in Prediction Models. In E.W. Steyerberg, editor, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, Statistics for Biology and Health, pages 83–100. Springer, New York, NY, 2009.

[18] Stephen Jay Gould. The Median Isn't the Message. *AMA Journal of Ethics*, 15(1):77–81, January 2013.

[19] Morteza Abdullatif Khafaie and Fakher Rahim. Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2. *Osong Public Health and Research Perspectives*, 11(2):74–80, April 2020.

[20] Google. Uk current death rate covid - Google Search. https://www.google.com/search?q=uk+current+death+rate+covid&rlz=1C1GCEU_en-GBGB932GB935&oq=uk+current+death&aqs=chrome.1.0i512l2j69i57j0i512j0i22i30l3j0i10i22i30j69i64.3 8, 2021.

[21] UK Government. UK Summary — Coronavirus (COVID-19) in the UK. https://coronavirus.data.gov.uk, 2021.

[22] NICE. NICE recommends wider use of statins for prevention of CVD — News and features — News. https://www.nice.org.uk/news/article/nice-recommends-wider-use-of-statins-for-prevention-of-cvd, June 2014.

[23] Tony S. Mok, Yi-Long Wu, Sumitra Thongprasert, Chih-Hsin Yang, Da-Tong Chu, Nagahiro Saijo, Patrapim Sunpaweravong, Baohui Han, Benjamin Margono, Yukito Ichinose, Yutaka Nishiwaki, Yuichiro Ohe, Jin-Ji Yang, Busyamas Chewaskulyong, Haiyi Jiang, Emma L. Duffield, Claire L. Watkins, Alison A. Armour, and Masahiro Fukuoka. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *The New England Journal of Medicine*, 361(10):947–957, September 2009.

[24] S. Mallal, D. Nolan, C. Witt, G. Masel, A. M. Martin, C. Moore, D. Sayer, A. Castley, C. Mamotte, D. Maxwell, I. James, and F. T. Christiansen. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet (London, England)*, 359(9308):727–732, March 2002.

[25] D. A. Rizzi and S. A. Pedersen. Causality in medicine: Towards a theory and terminology. *Theoretical Medicine*, 13(3):233–254, September 1992.

[26] Richard Doll. Uncovering the effects of smoking: Historical perspective. *Statistical Methods in Medical Research*, 7(2):87–117, April 1998.

[27] Tyler J. VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196–220, February 2013.

[28] J. L. Haybittle, R. W. Blamey, C. W. Elston, J. Johnson, P. J. Doyle, F. C. Campbell, R. I. Nicholson, and K. Griffiths. A Prognostic Index in Primary Breast Cancer. *British Journal of Cancer*, 45(3):361–366, March 1982.

[29] Tyler Vigen. *Spurious Correlations*. June 2017.

[30] Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea. *Statistics in Medicine*, 25(1):127–141, January 2006.

[31] Gary S. Collins and Douglas G. Altman. Predicting the 10 Year Risk of Cardiovascular Disease in the United Kingdom: Independent and External Validation of an Updated Version of QRISK2. *BMJ*, 344, June 2012.

[32] International Warfarin Pharmacogenetics Consortium, T. E. Klein, R. B. Altman, N. Eriksson, B. F. Gage, S. E. Kimmel, M.-T. M. Lee, N. A. Limdi, D. Page, D. M. Roden, M. J. Wagner, M. D. Caldwell, and J. A. Johnson. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *The New England Journal of Medicine*, 360(8):753–764, February 2009.

[33] Andrew J. Vickers, Ethan Basch, and Michael W. Kattan. Against Diagnosis. *Annals of Internal Medicine*, 149(3):200–203, August 2008.

[34] Danilo Fliser, Barbara Kollerits, Ulrich Neyer, Donna P. Ankerst, Karl Lhotta, Arno Lingenhel, Eberhard Ritz, and Florian Kronenberg. Fibroblast Growth Factor 23 (FGF23) Predicts Progression of Chronic Kidney Disease: The Mild to Moderate Kidney Disease (MMKD) Study. *Journal of the American Society of Nephrology*, 18(9):2600–2608, September 2007.

[35] M. G. Hunink, L. Goldman, A. N. Tosteson, M. A. Mittleman, P. A. Goldman, L. W. Williams, J. Tsevat, and M. C. Weinstein. The Recent Decline in Mortality from Coronary Heart Disease, 1980-1990. The Effect of Secular Trends in Risk Factors and Treatment. *JAMA*, 277(7):535–542, February 1997.

[36] D. D. Gladman, M. B. Urowitz, C. H. Goldsmith, P. Fortin, E. Ginzler, C. Gordon, J. G. Hanly, D. A. Isenberg, K. Kalunian, O. Nived, M. Petri, J. Sanchez-Guerrero, M. Snaith, and G. Sturfelt. The Reliability of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index in Patients with Systemic Lupus Erythematosus. *Arthritis and Rheumatism*, 40(5):809–813, May 1997.

[37] Ian N. Bruce, Aidan G. O'Keeffe, Vern Farewell, John G. Hanly, Susan Manzi, Li Su, Dafna D. Gladman, Sang-Cheol Bae, Jorge Sanchez-Guerrero, Juanita Romero-Diaz, Caroline Gordon, Daniel J. Wallace, Ann E. Clarke, Sasha Bernatsky, Ellen M. Ginzler, David A. Isenberg, Anisur Rahman, Joan T. Merrill, Graciela S. Alarcón, Barri J. Fessler, Paul R. Fortin, Michelle Petri, Kristjan Steinsson, Mary Anne Dooley, Munther A. Khamashta, Rosalind Ramsey-Goldman, Asad A. Zoma, Gunnar K. Sturfelt, Ola Nived, Cynthia Aranow, Meggan Mackay, Manuel Ramos-Casals, Ronald F. van Vollenhoven, Kenneth C. Kalunian, Guillermo Ruiz-Irastorza, Sam Lim, Diane L. Kamen, Christine A. Peschken, Murat Inanc, and Murray B. Urowitz. Factors Associated with Damage Accrual in Patients with Systemic Lupus Erythematosus: Results from the Systemic Lupus International Collaborating Clinics (SLICC) Inception Cohort. *Annals of the Rheumatic Diseases*, 74(9):1706–1713, September 2015.

[38] Nish Chaturvedi. ETHNIC DIFFERENCES IN CARDIOVASCULAR DISEASE. *Heart*, 89(6):681–686, June 2003.

[39] Paramjit S Gill, Gill Plumridge, Kamlesh Khunti, and Sheila Greenfield. Under-Representation of Minority Ethnic Groups in Cardiovascular Research: A Semi-Structured Interview Study. *Family Practice*, 30(2):233–241, April 2013.

[40] Manjula Kurella Tamura, Jane C. Tan, and Ann M. O'Hare. Optimizing Renal Replacement Therapy in Older Adults: A Framework for Making Individualized Decisions. *Kidney International*, 82(3):261–269, August 2012.

[41] M. Justin Zaman, Justin Zaman, and Eric Brunner. Social Inequalities and Cardiovascular Disease in South Asians. *Heart (British Cardiac Society)*, 94(4):406–407, April 2008.

[42] Stephen B. Hanauer. Exploring the Controversial Themes of IBD. *Inflammatory Bowel Diseases*, 15(S1):S1–S10, 2009.

[43] Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Rubin Minhas, Aziz Sheikh, and Peter Brindle. Predicting Cardiovascular Risk in England and Wales: Prospective Derivation and Validation of QRISK2. *BMJ*, 336(7659):1475–1482, June 2008.

[44] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford. Importance of Events per Independent Variable in Proportional Hazards Regression Analysis. II. Accuracy and Precision of Regression Estimates. *Journal of Clinical Epidemiology*, 48(12):1503–1510, December 1995.

[45] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, December 1996.

[46] Richard D. Riley, Kym IE Snell, Joie Ensor, Danielle L. Burke, Frank E. Harrell Jr, Karel GM Moons, and Gary S. Collins. Minimum Sample Size for Developing a Multivariable Prediction Model: PART II - Binary and Time-to-Event Outcomes. *Statistics in Medicine*, 38(7):1276–1296, 2019.

[47] Willi Sauerbrei, Patrick Royston, and Harald Binder. Selection of Important Variables and Determination of Functional Form for Continuous Predictors in Multivariable Model Building. *Statistics in Medicine*, 26(30):5512–5528, 2007.

[48] Yvonne Vergouwe, Ewout W. Steyerberg, Marinus J. C. Eijkemans, and J. Dik F. Habbema. Substantial Effective Sample Sizes Were Required for External Validation Studies of Predictive Logistic Regression Models. *Journal of Clinical Epidemiology*, 58(5):475–483, May 2005.

[49] Gary S. Collins, Emmanuel O. Ogundimu, and Douglas G. Altman. Sample Size Considerations for the External Validation of a Multivariable Prognostic Model: A Resampling Study. *Statistics in Medicine*, 35(2):214–226, January 2016.

[50] C. Counsell and M. Dennis. Systematic Review of Prognostic Models in Patients with Acute Stroke. *Cerebrovascular Diseases (Basel, Switzerland)*, 12(3):159–170, 2001.

[51] J. Ivanov, M. A. Borger, T. E. David, G. Cohen, N. Walton, and C. D. Naylor. Predictive Accuracy Study: Comparing a Statistical Model to Clinicians' Estimates of Outcomes after Coronary Bypass Surgery. *The Annals of Thoracic Surgery*, 70(1):162–168, July 2000.

[52] Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Margaret May, and Peter Brindle. Derivation and Validation of QRISK, a New Cardiovascular Disease Risk Score for the United Kingdom: Prospective Open Cohort Study. *BMJ (Clinical research ed.)*, 335(7611):136, July 2007.

[53] Gary S. Collins and Douglas G. Altman. An Independent and External Validation of QRISK2 Cardiovascular Disease Risk Score: A Prospective Open Cohort Study. *BMJ*, 340, May 2010.

[54] J. H. Todd, C. Dowle, M. R. Williams, C. W. Elston, I. O. Ellis, C. P. Hinton, R. W. Blamey, and J. L. Haybittle. Confirmation of a Prognostic Index in Primary Breast Cancer. *British Journal of Cancer*, 56(4):489–492, October 1987.

[55] Benoit Liquet, Jean-François Timsit, and Virginie Rondeau. Investigating Hospital Heterogeneity with a Multi-State Frailty Model: Application to Nosocomial Pneumonia Disease in Intensive Care Units. *BMC medical research methodology*, 12:79, June 2012.

[56] Kym I. E. Snell, Harry Hua, Thomas P. A. Debray, Joie Ensor, Maxime P. Look, Karel G. M. Moons, and Richard D. Riley. Multivariate Meta-Analysis of Individual Participant Data Helped Externally Validate the Performance and Implementation of a Prediction Model. *Journal of Clinical Epidemiology*, 69:40–50, January 2016.

[57] P. Hougaard. Frailty Models for Survival Data. *Lifetime Data Analysis*, 1(3):255–273, 1995.

[58] S. W. Lagakos. Effects of Mismodelling and Mismeasuring Explanatory Variables on Tests of Their Association with a Response Variable. *Statistics in Medicine*, 7(1-2):257–274, 1988.

[59] L. H. J. Eberhart, A. M. Morin, D. Guber, F. J. Kretz, A. Schäuffelen, H. Treiber, H. Wulf, and G. Geldner. Applicability of Risk Scores for Postoperative Nausea and Vomiting in Adults to Paediatric Patients. *British Journal of Anaesthesia*, 93(3):386–392, September 2004.

[60] Julia Hippisley-Cox, Carol Coupland, John Robson, and Peter M. Brindle. Advantages of QRISK2 (2010): The Key Issue Is Ethnicity and Extent of Reallocation. *Heart*, 2011.

[61] R. D. Riley, K. R. Abrams, A. J. Sutton, P. C. Lambert, D. R. Jones, D. Heney, and S. A. Burchill. Reporting of Prognostic Markers: Current Problems and Development of Guidelines for Evidence-Based Practice in the Future. *British Journal of Cancer*, 88(8):1191–1198, April 2003.

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

[63] P. Royston and W. Sauerbrei. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables. *undefined*, 2008.

[64] Raymond Hubbard and R. Murray Lindsay. Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing:. *Theory & Psychology*, February 2008.

[65] Patrick Royston, G Ambler, and W Sauerbrei. The Use of Fractional Polynomials to Model Continuous Risk Variables in Epidemiology. *International journal of epidemiology*, 28:964–74, November 1999.

[66] Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K. Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P. Martin. Informative Presence and Observation in Routine Health Data: A Review of Methodology for Clinical Risk Prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, November 2020.

[67] Patrick Royston and Mahesh K. B. Parmar. Flexible Parametric Proportional-Hazards and Proportional-Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects. *Statistics in Medicine*, 21(15):2175–2197, August 2002.

[68] Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976.

[69] Frank Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[70] Julia M. Rohrer. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, 1(1):27–42, March 2018.

[71] OO Aalen, K Røysland, JM Gran, R Kouyos, and T Lange. Can We Believe the DAGs? a Comment on the Relationship between Causal DAGs and Mechanisms. *Statistical Methods in Medical Research*, 25(5):2294–2314, October 2016.

[72] Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.

[73] jakewestfall. Using Causal Graphs to Understand Missingness and How to Deal with It, August 2017.

[74] A. Rogier T. Donders, Geert J. M. G. van der Heijden, Theo Stijnen, and Karel G. M. Moons. Review: A Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, October 2006.

[75] Ian R White and Patrick Royston. Imputing Missing Covariate Values for the Cox Model. *Statistics in Medicine*, 28(15):1982–1998, July 2009.

[76] S. Goodlad. Science for Non-Scientists : An Examination of Objectives and Constraints in the Presentation of Science to Non-Specialists, 1973.

[77] Laura J. Bonnett, Kym I. E. Snell, Gary S. Collins, and Richard D. Riley. Guide to Presenting Clinical Prediction Models for Use in Clinical Settings. *BMJ*, 365:l737, April 2019.

[78] Uchida Kazutaka, Yoshimura Shinichi, Hiyama Nagayasu, Oki Yoshiharu, Matsumoto Tsuyoshi, Tokuda Ryo, Yamaura Ikuya, Saito Shin, Takeuchi Masataka, Shigeta Keigo, Araki Hayato, and Morimoto Takeshi. Clinical Prediction Rules to Classify Types of Stroke at Prehospital Stage. *Stroke*, 49(8):1820–1827, August 2018.

[79] Alvin H. Moss, Jesse Ganjoo, Sanjay Sharma, Julie Gansor, Sharon Senft, Barbara Weaner, Cheryl Dalton, Karen MacKay, Beth Pellegrino, Priya Anantharaman, and Rebecca Schmidt. Utility of the "Surprise" Question to Identify Dialysis Patients with High Mortality. *Clinical Journal of the American Society of Nephrology*, 3(5):1379–1384, September 2008.

[80] Alexander Pate, Richard Emsley, Darren M. Ashcroft, Benjamin Brown, and Tjeerd van Staa. The Uncertainty with Using Risk Prediction Models for Individual Decision Making: An Exemplar Cohort Study Examining the Prediction of Cardiovascular Disease in English Primary Care. *BMC Medicine*, 17(1):134, July 2019.

[81] Prachi Bhatnagar, Kremlin Wickramasinghe, Julianne Williams, Mike Rayner, and Nick Townsend. The Epidemiology of Cardiovascular Disease in the UK 2014. *Heart*, 101(15):1182–1189, August 2015.

[82] David A. Jenkins, Matthew Sperrin, Glen P. Martin, and Niels Peek. Dynamic Models to Predict Health Outcomes: Current Status and Methodological Challenges. *Diagnostic and Prognostic Research*, 2(1):23, December 2018.

[83] QRISK3.

[84] Susan Mallett, Patrick Royston, Susan Dutton, Rachel Waters, and Douglas G. Altman. Reporting Methods in Studies Developing Prognostic Models in Cancer: A Review. *BMC Medicine*, 8(1):20, March 2010.

[85] Walter Bouwmeester, Nicolaas P. A. Zuithoff, Susan Mallett, Mirjam I. Geerlings, Yvonne Vergouwe, Ewout W. Steyerberg, Douglas G. Altman, and Karel G. M. Moons. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLOS Medicine*, 9(5):e1001221, May 2012.

[86] Lingxiao Chen. Overview of Clinical Prediction Models. *Annals of Translational Medicine*, 8(4), February 2020.

[87] S. E. Bleeker, H. A. Moll, E. W. Steyerberg, A. R. T. Donders, G. Derksen-Lubsen, D. E. Grobbee, and K. G. M. Moons. External Validation Is Necessary in Prediction Research: A Clinical Example. *Journal of Clinical Epidemiology*, 56(9):826–832, September 2003.

[88] Brendan M. Reilly and Arthur T. Evans. Translating Clinical Research into Clinical Practice: Impact of Using Prediction Rules to Make Decisions. *Annals of Internal Medicine*, 144(3):201–209, February 2006.

[89] QResearch. About - QResearch, 2020.

[90] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS - HARRELL - 1996 - Statistics in Medicine - Wiley Online Library, 1996.

[91] W. J. Mackillop and C. F. Quirt. Measuring the Accuracy of Prognostic Judgments in Oncology. *Journal of Clinical Epidemiology*, 50(1):21–29, January 1997.

[92] Michael J. Pencina, Ralph B. D'Agostino, Ralph B. D'Agostino, and Ramachandran S. Vasan. Evaluating the Added Predictive Ability of a New Marker: From Area under the ROC Curve to Reclassification and Beyond. *Statistics in Medicine*, 27(2):157–172; discussion 207–212, January 2008.

[93] Y. Liao, D. L. McGee, R. S. Cooper, and M. B. Sutkowski. How Generalizable Are Coronary Risk Prediction Models? comparison of Framingham and Two National Cohorts. *American Heart Journal*, 137(5):837–845, May 1999.

[94] Paolo Fraccaro, Sabine van der Veer, Benjamin Brown, Mattia Prosperi, Donal O'Donoghue, Gary S. Collins, Iain Buchan, and Niels Peek. An External Validation of Models to Predict the Onset of Chronic Kidney Disease Using Population-Based Electronic Health Records from Salford, UK. *BMC medicine*, 14:104, July 2016.

[95] Peter C. Austin, Frank E. Harrell, and David van Klaveren. Graphical Calibration Curves and the Integrated Calibration Index (ICI) for Survival Models. *Statistics in Medicine*, 39(21):2714–2742, 2020.

[96] A. C. Justice, K. E. Covinsky, and J. A. Berlin. Assessing the Generalizability of Prognostic Information. *Annals of Internal Medicine*, 130(6):515–524, March 1999.

[97] J. André Knottnerus. Between Iatrotropic Stimulus and Interiatric Referral: The Domain of Primary Care Research. *Journal of Clinical Epidemiology*, 55(12):1201–1206, December 2002.

[98] Is Most Published Research Wrong?, August 2016.

[99] John P.A. Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2005.

[100] Glen P. Martin, Mamas A. Mamas, Niels Peek, Iain Buchan, and Matthew Sperrin. A Multiple-Model Generalisation of Updating Clinical Prediction Models. *Statistics in Medicine*, 37(8):1343–1358, 2018.

[101] Sonja Grill, Donna P. Ankerst, Mitchell H. Gail, Nilanjan Chatterjee, and Ruth M. Pfeiffer. Comparison of Approaches for Incorporating New Information into Existing Risk Prediction Models. *Statistics in Medicine*, 36(7):1134–1156, 2017.

[102] Laure Wynants, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M. J. Bonten, Darren L. Dahly, Johanna A. A. Damen, Thomas

P. A. Debray, Valentijn M. T. de Jong, Maarten De Vos, Paula Dhiman, Maria C. Haller, Michael O. Harhay, Liesbet Henckaerts, Pauline Heus, Nina Kreuzberger, Anna Lohmann, Kim Luijken, Jie Ma, Glen P. Martin, Constanza L. Andaur Navarro, Johannes B. Reitsma, Jamie C. Sergeant, Chunhu Shi, Nicole Skoetz, Luc J. M. Smits, Kym I. E. Snell, Matthew Sperrin, René Spijker, Ewout W. Steyerberg, Toshihiko Takada, Ioanna Tzoulaki, Sander M. J. van Kuijk, Florien S. van Royen, Jan Y. Verbakel, Christine Wallisch, Jack Wilkinson, Robert Wolff, Lotty Hooft, Karel G. M. Moons, and Maarten van Smeden. Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal. *BMJ*, 369:m1328, April 2020.

[103] NICE. Overview — Cardiovascular Disease: Risk Assessment and Reduction, Including Lipid Modification — Guidance — NICE.

[104] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys: 81*. Wiley-Interscience, Hoboken, N.J, 1st edition edition, May 2004.

[105] NICE. CVD Risk Assessment and Management, March 2019.

[106] Julia Hippisley-Cox and Carol Coupland. Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales: Prospective Derivation and External Validation of the QKidney Scores. *BMC family practice*, 11:49, June 2010.

[107] I. Stiell, G. Wells, A. Laupacis, R. Brison, R. Verbeek, K. Vandemheen, and C. D. Naylor. Multicentre Trial to Introduce the Ottawa Ankle Rules for Use of Radiography in Acute Ankle Injuries. Multicentre Ankle Rule Study Group. *BMJ (Clinical research ed.)*, 311(7005):594–597, September 1995.

[108] C. Cameron and C. D. Naylor. No Impact from Active Dissemination of the Ottawa Ankle Rules: Further Evidence of the Need for Local Implementation of Practice Guidelines. *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne*, 160(8):1165–1168, April 1999.

[109] Marion K. Campbell, Diana R. Elbourne, and Douglas G. Altman. CONSORT Statement: Extension to Cluster Randomised Trials. *BMJ*, 328(7441):702–708, March 2004.

[110] Jonathan C. Hill, David GT Whitehurst, Martyn Lewis, Stirling Bryan, Kate M. Dunn, Nadine E. Foster, Kika Konstantinou, Chris J. Main, Elizabeth Mason, Simon Somerville, Gail Sowden, Kanchan Vohora, and Elaine M. Hay. Comparison of Stratified Primary Care Management for Low Back Pain with Current Best Practice (STarT Back): A Randomised Controlled Trial. *The Lancet*, 378(9802):1560–1571, October 2011.

[111] Noura Anwar and Mahmoud Riad Mahmoud. A Stochastic Model for the Progression of Chronic Kidney Disease. 4(11):12, 2014.

[112] Per Kragh Andersen and Niels Keiding. Multi-State Models for Event History Analysis. *Statistical Methods in Medical Research*, 11(2):91–115, April 2002.

[113] H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in Biostatistics: Competing Risks and Multi-State Models. *Statistics in Medicine*, 26(11):2389–2430, May 2007.

[114] Leo Breiman. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science*, 16(3):199–231, August 2001.

[115] D. Y. Lin. On the Breslow estimator. *Lifetime Data Analysis*, 13(4):471–480, December 2007.

[116] Terry M. Therneau and Patricia M. Grambsch. The Cox Model. In Terry M. Therneau and Patricia M. Grambsch, editors, *Modeling Survival Data: Extending the Cox Model*, Statistics for Biology and Health, pages 39–77. Springer, New York, NY, 2000.

[117] Sky McKinley and Megan Levine. Cubic Spline Interpolation. page 15.

[118] Per Kragh Andersen, Steen Z. Abildstrom, and Susanne Rosthøj. Competing Risks as a Multi-State Model. *Statistical Methods in Medical Research*, 11(2):203–215, April 2002.

[119] Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[120] David Cox. Partial Likelihood. *Biometrika*, 62:269–276, August 1975.

[121] DANIEL COMMENGES, PIERRE JOLY, ANNE GÉGOUT-PETIT, and BENOIT LIQUET. Choice between Semi-Parametric Estimators of Markov and Non-Markov Multi-State Models from Coarsened Observations. *Scandinavian Journal of Statistics*, 34(1):33–52, 2007.

[122] D. D. Gladman, C. H. Goldsmith, M. B. Urowitz, P. Bacon, P. Fortin, E. Ginzler, C. Gordon, J. G. Hanly, D. A. Isenberg, M. Petri, O. Nived, M. Snaith, and G. Sturfelt. The Systemic Lupus International Collaborating Clinics/American College of Rheumatology (SLICC/ACR) Damage Index for Systemic Lupus Erythematosus International Comparison. *The Journal of Rheumatology*, 27(2):373–376, February 2000.

[123] D. Commenges. Inference for Multi-State Models from Interval-Censored Data. *Statistical Methods in Medical Research*, 11(2):167–182, April 2002.

[124] Robert C. Atkins. The Epidemiology of Chronic Kidney Disease. *Kidney International*, 67:S14–S18, April 2005.

[125] Katherine T. Mills, Yu Xu, Weidong Zhang, Joshua D. Bundy, Chung-Shiuan Chen, Tanika N. Kelly, Jing Chen, and Jiang He. A Systematic Analysis of Worldwide Population-Based Data on the Global Burden of Chronic Kidney Disease in 2010. *Kidney International*, 88(5):950–957, November 2015.

[126] Julie Gilg, Fergus Caskey, and Damian Fogarty. UK Renal Registry 18th Annual Report: Chapter 1 UK Renal Replacement Therapy Incidence in 2014: National and Centre-Specific Analyses. *Nephron*, 132(Suppl. 1):9–40, 2016.

[127] David Ansell, TG Feest, C Tomson, AJ Williams, and G Warwick. 9th Annual Report of the Renal Association. *Nephrology Dialysis Transplantation*, 2007.

[128] Allan J. Collins, Robert N. Foley, Charles Herzog, Blanche Chavers, David Gilbertson, Areef Ishani, Bertram Kasiske, Jiannong Liu, Lih-Wen Mau, Marshall McBean, Anne Murray, Wendy St Peter, Haifeng Guo, Sally Gustafson, Qi Li, ShuLing Li, Suying Li, Yi Peng, Yang Qiu, Tricia Roberts, Melissa Skeans, Jon Snyder, Craig Solid, Changchun Wang, Eric Weinhandl, David Zaun, Cheryl Arko, Shu-Cheng Chen, Frederick Dalleska, Frank Daniels, Stephan Dunning, James Ebben, Eric Frazier, Christopher Hanzlik, Roger Johnson, Daniel Sheets, Xinyue Wang, Beth Forrest, Edward Constantini, Susan Everson, Paul Eggers, and Lawrence Agodoa. US Renal Data System 2010 Annual Data Report. *American Journal of Kidney Diseases*, 57(1):A8, January 2011.

[129] Ann M. O'Hare, Rudolph A. Rodriguez, Susan M. Hailpern, Eric B. Larson, and Manjula Kurella Tamura. Regional Variation in Health Care Intensity and Treatment Practices for End-Stage Renal Disease in Older Adults. *JAMA*, 304(2):180–186, July 2010.

[130] Ya-Chen Tina Shih, A. M. Y. Guo, Paul M. Just, and Salim Mujais. Impact of Initial Dialysis Modality and Modality Switches on Medicare Expenditures of End-Stage Renal Disease Patients. *Kidney International*, 68(1):319–329, July 2005.

[131] Robert N. Foley, Shu-Cheng CHEN, and Allan J. Collins. Hemodialysis Access at Initiation in the United States, 2005 to 2007: Still "Catheter First". *Hemodialysis International*, 13(4):533–542, 2009.

[132] Vascular Access Work Group. Clinical Practice Guidelines for Vascular Access. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 48 Suppl 1:S248–273, July 2006.

[133] Albert I. Richardson, Andrew Leake, Gregory C. Schmieder, Andre Biuckians, Gordon K. Stokes, Jean M. Panneton, and Marc H. Glickman. Should Fistulas Really Be First in the Elderly Patient? *The journal of vascular access*, 10(3):199–202, 2009.

[134] Jeffrey Perl, Ron Wald, Philip McFarlane, Joanne M. Bargman, Edward Vonesh, Yingbo Na, S. Vanita Jassal, and Louise Moist. Hemodialysis Vascular Access Modifies the Association between Dialysis Modality and Survival. *Journal of the American Society of Nephrology*, 22(6):1113–1121, 2011.

[135] Rajnish Mehrotra, Yi-Wen Chiu, Kamyar Kalantar-Zadeh, Joanne Bargman, and Edward Vonesh. Similar Outcomes with Hemodialysis and Peritoneal Dialysis in Patients with End-Stage Renal Disease. *Archives of internal medicine*, 171(2):110–118, 2011.

[136] Kidney Chain - UCLA Kidney Exchange Program - Los Angeles, CA.

[137] M Cooper and CL Forland. The Elderly as Recipients of Living Donor Kidneys, How Old Is Too Old? - Abstract - Europe PMC. *Current Opinion in Organ Transplantation*, 2011.

[138] Robert A. Wolfe, Valarie B. Ashby, Edgar L. Milford, Akinlolu O. Ojo, Robert E. Ettenger, Lawrence Y.C. Agodoa, Philip J. Held, and Friedrich K. Port. Comparison of Mortality in All Patients on Dialysis, Patients on Dialysis Awaiting Transplantation, and Recipients of a First Cadaveric Transplant. *New England Journal of Medicine*, 341(23):1725–1730, December 1999.

[139] Robert M. Merion, Valarie B. Ashby, Robert A. Wolfe, Dale A. Distant, Tempie E. Hulbert-Shearon, Robert A. Metzger, Akinlolu O. Ojo, and Friedrich K. Port. Deceased-Donor Characteristics and the Survival Benefit of Kidney Transplantation. *JAMA*, 294(21):2726–2733, December 2005.

[140] Morgan E. Grams, Yingying Sang, Shoshana H. Ballew, Juan Jesus Carrero, Ognjenka Djurdjev, Hiddo J. L. Heerspink, Kevin Ho, Sadayoshi Ito, Angharad Marks, David Naimark, Danielle M. Nash, Sankar D. Navaneethan, Mark Sarnak, Benedicte Stengel, Frank L. J. Visseren, Angela Yee-Moon Wang, Anna Köttgen, Andrew S. Levey, Mark Woodward, Kai-Uwe Eckardt, Brenda Hemmelgarn, and Josef Coresh. Predicting Timing of Clinical Outcomes in Patients with Chronic Kidney Disease and Severely Decreased Glomerular Filtration Rate. *Kidney International*, 93(6):1442–1451, June 2018.

[141] A. Begun, A. Icks, R. Waldeyer, S. Landwehr, M. Koch, and G. Giani. Identification of a Multistate Continuous-Time Nonhomogeneous Markov Chain Model for Patients with Decreased Renal Function. *Medical decision making : an international journal of the Society for Medical Decision Making*, 33(2):298–306, February 2013.

[142] K. J. Jager, V. S. Stel, C. Wanner, C. Zoccali, and F. W. Dekker. The Valuable Contribution of Observational Studies to Nephrology. *Kidney International*, 72(6):671–675, September 2007.

[143] P. M. Rothwell. External Validity of Randomised Controlled Trials: 'to Whom Do the Results of This Trial Apply? *Lancet*, 365, 2005.

[144] Z. Fewell, G. Davey Smith, and J. A. C. Sterne. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *Am J Epidemiol*, 166, 2007.

[145] N. X. Lin, S. Logan, and W. E. Henley. Bias and Sensitivity Analysis When Estimating Treatment Effects from the Cox Model with Omitted Covariates. *Biometrics*, 69, 2013.

[146] G. K. Reeves, K. Pirie, V. Beral, J. Green, E. Spencer, and D. Bull. Cancer Incidence and Mortality in Relation to Body Mass Index in the Million Women Study: Cohort Study. *BMJ*, 335, 2007.

[147] O. Klungsøyr, J. Sexton, I. Sandanger, and J. F. Nygård. Sensitivity Analysis for Unmeasured Confounding in a Marginal Structural Cox Proportional Hazards Model. *Lifetime Data Anal*, 15, 2009.

[148] W. Chen, X. Zhang, D. E. Faries, W. Shen, J. W. Seaman, and J. D. Stamey. A Bayesian Approach to Correct for Unmeasured or Semi-Unmeasured Confounding in Survival Data Using Multiple Validation Data Sets. *Epidemiol Biostat Public Heal*, 14, 2017.

[149] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLoS Med*, 4, 2007.

[150] K. B. Pouwels, N. N. Widyakusuma, R. H. H. Groenwold, and E. Hak. Quality of Reporting of Confounding Remained Suboptimal after the STROBE Guideline. *J Clin Epidemiol*, 69, 2016.

[151] T. J. Vanderweele and P. Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med Ann*, 1677326, 2017.

[152] Michael T. Koller, Heike Raatz, Ewout W. Steyerberg, and Marcel Wolbers. Competing Risks and the Clinical Community: Irrelevance or Ignorance? *Statistics in Medicine*, 31(11-12):1089–1097, May 2012.

[153] D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall Ltd, Cambridge, 1984.

[154] Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. The Design of Simulation Studies in Medical Statistics. *Statistics in Medicine*, 25(24):4279–4292, December 2006.

[155] M. J. Crowther and P. C. Lambert. Simulating Biologically Plausible Complex Survival Data. *Stat Med*, 32, 2013.

[156] B. Haller and K. Ulm. Flexible Simulation of Competing Risks Data Following Prespecified Subdistribution Hazards. *J Stat Comput Simul*, 84, 2014.

[157] whuber. Generate a Gaussian and a Binary Random Variables with Predefined Correlation, November 2017.

[158] N. Grambauer, M. Schumacher, and J. Beyersmann. Proportional Subdistribution Hazards Modeling Offers a Summary Analysis, Even If Misspecified. *Stat Med*, 29, 2010.

[159] A Latouche, V Boisson, S Chevret, and R Porcher. Misspecified Regression Model for the Subdistribution Hazard of a Competing Risk. *Statistics in Medicine*, 26(5):965–974, February 2007.

[160] M. H. Gail, S. Wieand, and S. Piantadosi. Biased Estimates of Treatment Effects in Randomized Experiments with Nonlinear Regression and Omitted Covariates. *Biometrika*, 71, 1984.

[161] F. Mosteller. A K-Sample Slippage Test for an Extreme Population. *Ann Math Stat*, 19, 1948.

[162] Caroline A. Thompson, Zuo-Feng Zhang, and Onyebuchi A. Arah. Competing Risk Bias to Explain the Inverse Relationship between Smoking and Malignant Melanoma. *European Journal of Epidemiology*, 28(7):557–567, July 2013.

[163] F. Song, A. A. Qureshi, X. Gao, T. Li, and J. Han. Smoking and Risk of Skin Cancer: A Prospective Analysis and a Meta-Analysis. *Int J Epidemiol*, 41, 2012.

[164] R. o. l. f. H. H. Groenwold, D. a. v. i. d. B. Nelson, K. r. i. s. t. i. n. L. Nichol, A. r. n. o. W. Hoes, and E. e. l. k. o. Hak. Sensitivity Analyses to Estimate the Potential Impact of Unmeasured Confounding in Causal Research. *International Journal of Epidemiology*, 39, 2009.

[165] M. M. Suttorp, B. Siegerink, K. J. Jager, C. Zoccali, and F. W. Dekker. Graphical Presentation of Confounding in Directed Acyclic Graphs. *Nephrol Dial Transplant*, 30, 2015.

[166] D. Y. Lin and L. J. Wei. The Robust Inference for the Cox Proportional Hazards Model. *J Am Stat Assoc*, 84, 1989.

[167] Catherine R. Lesko and Bryan Lau. Bias Due to Confounders for the Exposure-Competing Risk Relationship. *Epidemiology (Cambridge, Mass.)*, 28(1):20–27, January 2017.

[168] Ewout W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer Science & Business Media, December 2008.

[169] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. A Calibration Hierarchy for Risk Models Was Defined: From Utopia to Empirical Data. *Journal of Clinical Epidemiology*, 74:167–176, June 2016.

[170] Cynthia S. Crowson, Elizabeth J. Atkinson, and Terry M. Therneau. Assessing Calibration of Prognostic Risk Scores. *Statistical methods in medical research*, 25(4):1692–1706, August 2016.

[171] Patrick Royston and Douglas G. Altman. External Validation of a Cox Prognostic Model: Principles and Methods. *BMC Medical Research Methodology*, 13(1):33, March 2013.

[172] David C. Goff, Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. O'Donnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Paul Sorlie, Neil J. Stone, and Peter W. F. Wilson. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*, 129(25 suppl 2):S49–S73, June 2014.

[173] B. de la Iglesia, J. F. Potter, N. R. Poulter, M. M. Robins, and J. Skinner. Performance of the ASSIGN Cardiovascular Disease Risk Score on a UK Cohort of Patients from General Practice. *Heart*, 97(6):491–499, March 2011.

[174] Patrick Royston. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities:. *The Stata Journal*, December 2014.

[175] Patrick Royston. Tools for Checking Calibration of a Cox Model in External Validation: Prediction of Population-Averaged Survival Curves Based on Risk Groups. *The Stata Journal*, 15(1):275–291, April 2015.

[176] Maja Pohar Perme and Per Kragh Andersen. Checking Hazard Regression Models Using Pseudo-Observations. *Statistics in medicine*, 27(25):5309–5328, November 2008.

[177] Thomas A. Gerds and Martin Schumacher. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, 48(6):1029–1040, 2006.

[178] Cristian Spitoni, Violette Lammens, and Hein Putter. Prediction Errors for State Occupation and Transition Probabilities in Multi-State Models. *Biometrical Journal. Biometrische Zeitschrift*, 60(1):34–48, January 2018.

[179] Xiaoxia Han, Yilong Zhang, and Yongzhao Shao. On Comparing Two Correlated C Indices with Censored Survival Data. *Statistics in medicine*, 36(25):4041–4049, November 2017.

[180] Xinhua Liu, Zhezhen Jin, and Joseph H. Graziano. Comparing Paired Biomarkers in Predicting Quantitative Health Outcome Subject to Random Censoring. *Statistical methods in medical research*, 25(1):447–457, February 2016.

[181] Xu Shu and Douglas E. Schaubel. Methods for Contrasting Gap Time Hazard Functions: Application to Repeat Liver Transplantation. *Statistics in biosciences*, 9(2):470–488, December 2017.

[182] Per Kragh Andersen and Maja Pohar Perme. Pseudo-Observations in Survival Analysis. *Statistical Methods in Medical Research*, 19(1):71–99, February 2010.

[183] R. Core Team. R: A Language and Environment for Statistical Computing, February 2020.

[184] Hadley Wickham. The Tidy Tools Manifesto, November 2017.

[185] Terry Therneau. A Package for Survival Analysis in R. page 89, March 2020.

[186] Maja Pohar Perme, Mette Gerster, and Kevin Rodrigues. Pseudo: Computes Pseudo-Observations for Modeling, July 2017.

[187] Dirk Eddelbuettel and Romain Francois. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(1):1–18, April 2011.

[188] Winston Chang, Joe Cheng, J. J. Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation (jQuery library and jQuery UI library), jQuery contributors (jQuery library; authors listed in inst/www/shared/jquery-AUTHORS.txt), jQuery UI contributors (jQuery UI library; authors listed in inst/www/shared/jqueryui/AUTHORS.txt), Mark Otto (Bootstrap library), Jacob Thornton (Bootstrap library), Bootstrap contributors (Bootstrap library), Twitter, Inc (Bootstrap library), Alexander Farkas (html5shiv library), Scott Jehl (Respond js library), Stefan Petre (Bootstrap-datepicker library),

Andrew Rowls (Bootstrap-datepicker library), Dave Gandy (Font-Awesome font), Brian Reavis (selectize js library), Kristopher Michael Kowal (es5-shim library), es5-shim contributors (es5-shim library), Denis Ineshin (ion rangeSlider library), Sami Samhuri (Javascript strftime library), SpryMedia Limited (DataTables library), John Fraser (showdown js library), John Gruber (showdown js library), Ivan Sagalaev (highlight js library), and R. Core Team (tar implementation from R). Shiny: Web Application Framework for R, March 2020.

[189] D. R. Cox. Two Further Applications of a Model for Binary Regression. 1958.

[190] Ewout W. Steyerberg and Yvonne Vergouwe. Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation. *European Heart Journal*, 35(29):1925–1931, August 2014.

[191] Glenn W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, January 1950.

[192] Kirsten Van Hoorde, Yvonne Vergouwe, Dirk Timmerman, Sabine Van Huffel, Ewout W. Steyerberg, and Ben Van Calster. Assessing Calibration of Multinomial Risk Prediction Models. *Statistics in Medicine*, 33(15):2585–2596, July 2014.

[193] Ben Van Calster, Vanya Van Belle, Yvonne Vergouwe, Dirk Timmerman, Sabine Van Huffel, and Ewout W. Steyerberg. Extending the C-Statistic to Nominal Polytomous Outcomes: The Polytomous Discrimination Index. *Statistics in Medicine*, 31(23):2610–2626, 2012.

[194] M. Schumacher, E. Graf, and T. Gerds. How to Assess Prognostic Models for Survival Data: A Case Study in Oncology. *Methods of Information in Medicine*, 42(5):564–571, 2003.

[195] Yutaka Matsuyama and Takuhiro Yamaguchi. Estimation of the Marginal Survival Time in the Presence of Dependent Competing Risks Using Inverse Probability of Censoring Weighted (IPCW) Methods. *Pharmaceutical Statistics*, 7(3):202–214, 2008.

[196] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers*. Wiley, New York, 3rd ed edition, 2003.

[197] A. Allen Bradley, Stuart S. Schwartz, and Tempei Hashino. Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score. *Weather and Forecasting*, 23(5):992–1006, October 2008.

[198] Peter C. Austin and Ewout W. Steyerberg. Interpreting the Concordance Statistic of a Logistic Regression Model: Relation to the Variance and Odds Ratio of a Continuous Explanatory Variable. *BMC Medical Research Methodology*, 12(1):82, June 2012.

[199] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988.

[200] Mario A. Cleves. Comparative Assessment of Three Common Algorithms for Estimating the Variance of the Area under the Nonparametric Receiver Operating Characteristic Curve. *The Stata Journal*, 2(3):280–289, September 2002.

[201] Valentijn M. T. de Jong, Marinus J. C. Eijkemans, Ben van Calster, Dirk Timmerman, Karel G. M. Moons, Ewout W. Steyerberg, and Maarten van Smeden. Sample Size Considerations and Predictive Performance of Multinomial Logistic Prediction Models. *Statistics in Medicine*, 38(9):1601–1619, 2019.

[202] Brian Ripley and William Venables. Package 'Nnet'. February 2016.

[203] Kenneth J. Arrow. *Social Choice and Individual Values: Third Edition.* Yale University Press, June 2012.

[204] K. Van Hoorde, S. Van Huffel, D. Timmerman, T. Bourne, and B. Van Calster. A Spline-Based Tool to Assess and Visualize the Calibration of Multiclass Risk Predictions. *Journal of Biomedical Informatics*, 54:283–293, April 2015.

[205] Eric S. Johnson, Micah L. Thorp, Xiuhai Yang, Olivier L. Charansonney, and David H. Smith. Predicting Renal Replacement Therapy and Mortality in CKD. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 50(4):559–565, October 2007.

[206] James P. Wick, Tanvir C. Turin, Peter D. Faris, Jennifer M. MacRae, Robert G. Weaver, Marcello Tonelli, Braden J. Manns, and Brenda R. Hemmelgarn. A Clinical Risk Prediction Tool for 6-Month Mortality After Dialysis Initiation Among Older Adults. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 69(5):568–575, May 2017.

[207] Martin J. Landray, Jonathan R. Emberson, Lisa Blackwell, Tanaji Dasgupta, Rosita Zakeri, Matthew D. Morgan, Charlie J. Ferro, Susan Vickery, Puja Ayrton, Devaki Nair, R. Neil Dalton, Edmund J. Lamb, Colin Baigent, Jonathan N. Townend, and David C. Wheeler. Prediction of ESRD and Death among People with CKD: The Chronic Renal Impairment in Birmingham (CRIB) Prospective Cohort Study. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 56(6):1082–1094, December 2010.

[208] Angharad Marks, Nicholas Fluck, Gordon J. Prescott, Lynn Robertson, William G. Simpson, William Cairns Smith, and Corri Black. Looking to the Future: Predicting Renal Replacement Outcomes in a Large Community Cohort with Chronic Kidney Disease. *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association*, 30(9):1507–1517, September 2015.

[209] Eric S. Johnson, Micah L. Thorp, Robert W. Platt, and David H. Smith. Predicting the Risk of Dialysis and Transplant among Patients with CKD: A Retrospective Cohort Study. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 52(4):653–660, October 2008.

[210] Sanjay Kulkarni, Isaac Hall, Richard Formica, Carrie Thiessen, Darren Stewart, Geliang Gan, Erich Greene, and Yanhong Deng. Transition Probabilities between Changing Sensitization Levels, Waitlist Activity Status and Competing-Risk Kidney Transplant Outcomes Using Multi-State Modeling. *PLOS ONE*, 12(12):e0190277, December 2017.

[211] Jürgen Floege, Iain A. Gillespie, Florian Kronenberg, Stefan D. Anker, Ioanna Gioni, Sharon Richards, Ronald L. Pisoni, Bruce M. Robinson, Daniele Marcelli, Marc Froissart, and Kai-Uwe Eckardt. Development and Validation of a Predictive Mortality Risk Score from a European Hemodialysis Cohort. *Kidney International*, 87(5):996–1008, May 2015.

[212] Xue-Ying Cao, Jian-Hui Zhou, Guang-Yan Cai, Ni-Na Tan, Jing Huang, Xiang-Cheng Xie, Li Tang, and Xiang-Mei Chen. Predicting One-Year Mortality in Peritoneal Dialysis Patients: An Analysis of the China Peritoneal Dialysis Registry. *International Journal of Medical Sciences*, 12(4):354–361, 2015.

[213] Navdeep Tangri, Lesley A. Stevens, John Griffith, Hocine Tighiouart, Ognjenka Djurdjev, David Naimark, Adeera Levin, and Andrew S. Levey. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA*, 305(15):1553–1559, April 2011.

[214] Navdeep Tangri, Lesley A. Inker, Brett Hiebert, Jenna Wong, David Naimark, David Kent, and Andrew S. Levey. A Dynamic Predictive Model for Progression of CKD. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 69(4):514–520, April 2017.

[215] Michael G. Shlipak, Linda F. Fried, Mary Cushman, Teri A. Manolio, Do Peterson, Catherine Stehman-Breen, Anthony Bleyer, Anne Newman, David Siscovick, and Bruce Psaty. Cardiovascular Mortality Risk in Chronic Kidney Disease: Comparison of Traditional and Novel Risk Factors. *JAMA*, 293(14):1737–1745, April 2005.

[216] John J. V. McMurray, Hajime Uno, Petr Jarolim, Akshay S. Desai, Dick de Zeeuw, Kai-Uwe Eckardt, Peter Ivanovich, Andrew S. Levey, Eldrin F. Lewis, Janet B. McGill, Patrick Parfrey, Hans-Henrik Parving, Robert M. Toto, Scott D. Solomon, and Marc A. Pfeffer. Predictors of Fatal and Nonfatal Cardiovascular Events in Patients with Type 2 Diabetes Mellitus, Chronic Kidney Disease, and Anemia: An Analysis of the Trial to Reduce Cardiovascular Events with Aranesp (Darbepoetin-Alfa) Therapy (TREAT). *American Heart Journal*, 162(4):748–755.e3, October 2011.

[217] Morgan E. Grams and Josef Coresh. Assessing Risk in Chronic Kidney Disease: A Methodological Review. *Nature Reviews. Nephrology*, 9(1):18–25, January 2013.

[218] Navdeep Tangri, Georgios D. Kitsios, Lesley Ann Inker, John Griffith, David M. Naimark, Simon Walker, Claudio Rigatto, Katrin Uhlig, David M. Kent, and Andrew S. Levey. Risk Prediction Models for Patients with Chronic Kidney Disease: A Systematic Review. *Annals of Internal Medicine*, 158(8):596–603, April 2013.

[219] Chava L. Ramspek, Pauline Wm Voskamp, Frans J. van Ittersum, Raymond T. Krediet, Friedo W. Dekker, and Merel van Diepen. Prediction Models for the Mortality Risk in

Chronic Dialysis Patients: A Systematic Review and Independent External Validation Study. *Clinical Epidemiology*, 9:451–464, 2017.

[220] Nisha Bansal, Ronit Katz, Ian H. De Boer, Carmen A. Peralta, Linda F. Fried, David S. Siscovick, Dena E. Rifkin, Calvin Hirsch, Steven R. Cummings, Tamara B. Harris, Stephen B. Kritchevsky, Mark J. Sarnak, Michael G. Shlipak, and Joachim H. Ix. Development and Validation of a Model to Predict 5-Year Risk of Death without ESRD among Older Adults with CKD. *Clinical journal of the American Society of Nephrology: CJASN*, 10(3):363–371, March 2015.

[221] Adler Perotte, Rajesh Ranganath, Jamie S. Hirsch, David Blei, and Noémie Elhadad. Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis. *Journal of the American Medical Informatics Association: JAMIA*, 22(4):872–880, July 2015.

[222] Alina M. Allen, W. Ray Kim, Terry M. Therneau, Joseph J. Larson, Julie K. Heimbach, and Andrew D. Rule. Chronic Kidney Disease and Associated Mortality after Liver Transplantation–a Time-Dependent Analysis Using Measured Glomerular Filtration Rate. *Journal of Hepatology*, 61(2):286–292, August 2014.

[223] Richard A. Hoefield, Philip A. Kalra, Patricia Baker, Beverley Lane, John P. New, Donal J. O'Donoghue, Robert N. Foley, and Rachel J. Middleton. Factors Associated with Kidney Disease Progression and Mortality in a Referred CKD Population. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, 56(6):1072–1081, December 2010.

[224] Rajkumar Chinnadurai, Constantina Chrysochou, and Philip A. Kalra. Increased Risk for Cardiovascular Events in Patients with Diabetic Kidney Disease and Non-Alcoholic Fatty Liver Disease. *Nephron*, 141(1):24–30, 2019.

[225] Andrew S. Levey, Lesley A. Stevens, Christopher H. Schmid, Yaping Lucy Zhang, Alejandro F. Castro, Harold I. Feldman, John W. Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, Josef Coresh, and CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A New Equation to Estimate Glomerular Filtration Rate. *Annals of Internal Medicine*, 150(9):604–612, May 2009.

[226] John P New, Nawar Diar Bakerly, David Leather, and Ashley Woodcock. Obtaining Real-World Evidence: The Salford Lung Study. *Thorax*, 69:1152–1154, 2014.

[227] Kunihiro Matsushita, Shoshana H. Ballew, Brad C. Astor, Paul E. de Jong, Ron T. Gansevoort, Brenda R. Hemmelgarn, Andrew S. Levey, Adeera Levin, Chi-Pang Wen, Mark Woodward, and Josef Coresh. Cohort Profile: The Chronic Kidney Disease Prognosis Consortium. *International Journal of Epidemiology*, 42(6):1660–1668, December 2013.

[228] Gopalakrishnan Venkat-Raman, Charles R.V. Tomson, Yongsheng Gao, Ronald Cornet,

Benedicte Stengel, Carola Gronhagen-Riska, Chris Reid, Christian Jacquelinet, Elke Schaeffner, Els Boeschoten, Francesco Casino, Frederic Collart, Johan De Meester, Oscar Zurriaga, Reinhard Kramar, Kitty J. Jager, and Keith Simpson. New Primary Renal Diagnosis Codes for the ERA-EDTA. *Nephrology Dialysis Transplantation*, 27(12):4414–4419, December 2012.

[229] Csaba P. Kovesdy, Josef Coresh, Shoshana H. Ballew, Mark Woodward, Adeera Levin, David M. J. Naimark, Joseph Nally, Dietrich Rothenbacher, Benedicte Stengel, Kunitoshi Iseki, Kunihiro Matsushita, and Andrew S. Levey. Past Decline Versus Current eGFR and Subsequent ESRD Risk. *Journal of the American Society of Nephrology*, 27(8):2447–2455, August 2016.

[230] David M. J. Naimark, Morgan E. Grams, Kunihiro Matsushita, Corri Black, Iefke Drion, Caroline S. Fox, Lesley A. Inker, Areef Ishani, Sun Ha Jee, Akihiko Kitamura, Janice P. Lea, Joseph Nally, Carmen Alicia Peralta, Dietrich Rothenbacher, Seungho Ryu, Marcello Tonelli, Hiroshi Yatsuya, Josef Coresh, Ron T. Gansevoort, David G. Warnock, Mark Woodward, and Paul E. de Jong. Past Decline Versus Current eGFR and Subsequent Mortality Risk. *Journal of the American Society of Nephrology*, 27(8):2456–2466, August 2016.

[231] Patricia M. Grambsch and Terry M. Therneau. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*, 81(3):515–526, 1994.

[232] David Schoenfeld. Partial Residuals for the Proportional Hazards Regression Model. *Biometrika*, 69(1):239–241, April 1982.

[233] Luís Meira-Machado, Jacobo de Una-Alvarez, Carmen Cadarso-Suarez, and Per K Andersen. Multi-State Models for the Analysis of Time-to-Event Data. *Statistical methods in medical research*, 18(2):195–222, April 2009.

[234] Kristel J. M. Janssen, Yvonne Vergouwe, A. Rogier T. Donders, Frank E. Harrell, Qingxia Chen, Diederick E. Grobbee, and Karel G. M. Moons. Dealing with Missing Predictor Values When Applying Clinical Prediction Models. *Clinical Chemistry*, 55(5):994–1001, May 2009.

[235] Angela M. Wood, Ian R. White, and Patrick Royston. How Should Variable Selection Be Performed with Multiply Imputed Data? *Statistics in Medicine*, 27(17):3227–3246, July 2008.

[236] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc, New York, NY, 1984.

[237] Michael Schomaker and Christian Heumann. Bootstrap Inference When Using Multiple Imputation. *Statistics in Medicine*, 37(14):2252–2266, 2018.

[238] Stef van Buuren. Package 'mice'. January 2021.

[239] Christopher Jackson. Flexsurv: A Platform for Parametric Survival Modelling in R. page 33, 2016.

[240] Davis Vaughan and Matt Dancho. Furrr: Apply Mapping Functions in Parallel Using Futures, May 2018.

[241] Stephen Jivraj and Omar Khan. Ethnicity and Deprivation in England: How Likely Are Ethnic Minorities to Live in Deprived Neighbourhoods? Technical report, University of Manchester, December 2013.

[242] Public Health England. Salford Unitary Authority Health Profile 2017. Technical report, July 2017.

[243] Karel G. M. Moons, Patrick Royston, Yvonne Vergouwe, Diederick E. Grobbee, and Douglas G. Altman. Prognosis and Prognostic Research: What, Why, and How? *BMJ*, 338:b375, February 2009.

[244] Ryan Ng, Kathy Kornas, Rinku Sutradhar, Walter P. Wodchis, and Laura C. Rosella. The Current Application of the Royston-Parmar Model for Prognostic Modeling in Health Research: A Scoping Review. *Diagnostic and Prognostic Research*, 2(1):4, December 2018.

[245] Glen P. Martin, David A. Jenkins, Lucy Bull, Rose Sisk, Lijing Lin, William Hulme, Anthony Wilson, Wenjuan Wang, Michael Barrowman, Camilla Sammut-Powell, Alexander Pate, Matthew Sperrin, and Niels Peek. Toward a Framework for the Design, Implementation, and Reporting of Methodology Scoping Reviews. *Journal of Clinical Epidemiology*, 127:191–197, 2020.

[246] Stef van Buuren. *Flexible Imputation of Missing Data*. second edition, 2018.

[247] Christopher Jackson. Multi-state modelling with R: The msm package. page 57, 2021.

[248] Liesbeth C. de Wreede, Marta Fiocco, and Hein Putter. Mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, 38(1):1–30, January 2011.

[249] Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.