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

2020

By
Michael Andrew Barrowman

Contents

List of Tables

List of Figures

Abstract

Insert Abstract Here...

Declaration

No portion of the work referred to in this 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.

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487`), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see `http://www.manchester.ac.uk/library/aboutus/regulations`) and in The University's policy on presentation of Theses

Acknowledgements

Insert Ackowledgements Here...

Introduction

Welcome to my Thesis. I've used R Markdown to create a gitbook style thesis, as well as a traditional pdf (following the UoM Thesis template). The html version can be found here.

I sent an email with details of how you can log in to the hypothes.is system. This will let you add comments to the gitbook pages. This is actually a really useful tool and has the ability to add annotations to any webpage. Annotated text can be seen highlighted and to add your own, just highlight the text you want to comment on, and the annotate bubble pops up (try it now). Make sure you click Post to Public after writing the annotation and annotations can include Rich Text and Markdown. You can view Annotations by clicking the menu on the right and this will show all comments.

Google Docs was great, but when I started to use R Markdown for other purposes, I realised I could embed tables and equations much easier and automatically using `kable`. Google Docs on it's own became rather clunky (especially with tables, just like Word) and whenever I updated a document using R Markdown, any comments that were added to the document were lost. It was also difficult for you to remember *where* each document lived. Since R Markdown can also output to `LaTeX` I thought it would be easier to collate papers together. I'd have had to do this kind of transition over to`LaTeX` for my thesis eventually and this would have been difficult and tedious from Google Docs or Word.

Importantly, I've also added in functionality to output each chapter individually. For unpublished papers, there will be a link to download the pdf and tex versions of the individual paper.

Everything here is hosted in a Github repo. I originally used Github as my backup, but then decided to increase how I used it.

Chapter 1

Literature Report

Last updated: 01 May

1.1 Introduction

1.2 Clinical Prediction Models

The idea of prognosis dates back to ancient Greece with the work of Hippocrates [1] and is derived from the Greek for "know before" meaning to forecast the future. Within the sphere of healthcare, it is definde 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 uaully given as a probability of having an event in a prespecified number of years. For example, QRISK3 [2] 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 modeling 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 ffect of screening for asymptomatic diseases (such as with mammograms[3]). Prognosis research can be used to develop new definitions of diseases, whether a redefinition of an existing disease (such as the extension to th definition of myocardial infarction to include non-fatal events [4]) or a previously unknown subtype of a disease (such as Brugada syndrome as a type of cardiovascular disease[5])

In general, prognosis research can be broken down into four main categories, with three subcategories [6]:

- Type I: Fundamental prognosis research [3]
- Type II: Prognostic factor research [7]
- Type III: Prognostic model research [8]
  - Model development [9]
  - Model validation [10]
  - Model impact evaluation [11]
- Type IV: Stratified Medicine [12]

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 [8]. External validation is considered by some to be more important than the actual deviration of the model as it demonstrates generalisability of the model [16], whereas a model on it's own may be highly susceptible to overfitting [**Cite: Something**].

1.2.1 Fundamental Prognosis Research

[**What is it? Old definition is incorrect, so will need to write this fresh**]

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.

Predictive 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 *can* affect an outcome, we cannot figure out which variables *will* affect the outcome. For example, [**xxxx**] demonstrated that [**xxxx**] is correlated with [**xxxx**], which subsequently used as a covariate in the development of the [**xxxx**] model. Note the use of the word correlate here as prognostic relationships do not have to be causal ones [**Cite: Something**]. These factors may indeed represent an underlying causal pathway, but this is not a requirement and it would require aetiological methods to discern whether it were causal or not. For example, when predicting [**xxxx**], we can demonstrate that [**xxxx**] is a prognostic factor, [however since the arrow of causation is [**xxxx**]] [**OR**] [however since [**xxxx**] causes both [**xxxx**] and [**xxxx**]], the relationship is prognostic, but not causal. [**Previously used Apgar score here, reference 40**]

Counter to the idea that prognostic factors aren't always causal, they are *always* confounding factors for the event they predict. Thue 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 [7]. 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 [17]. 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.

It is unfortunate, however, that Riley at al **???** found that 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 [7].

1.2.3 Prognostic Model Research

Predictive factors can be combined into a predictive model, which is a much more specific measurement of the effect of a factor on an outcome [8] and they are deigned to augment the job of a clinician; and not to completely replace them [11]. 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;~moons_transparent_2015]. 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 stratied 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 [11]. 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 **???**. 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.

Creating a clinicaly useful model is not as simple as just using some availble data to develop a model, despite what a lot of researchers seem to believe [**Cite: Something**]. To quote Steyerberg et al [8]. " To be useful for clinicia,s 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 mdel might appear to be useful, its effectiveness is only relevant to the population it was developed in. If your population is different, then the model will behave differently. Bleeker [18] developed a model to predict bacterial infections in febrile children with an unknown source. The model scored well when assessed for the predictive value in the development dataset, however it scored much worse in an external dataset implying that, though it worked well in the development population, it would be unwise to apply it to a new population.

Model Development

The first stage of having a useful model is to develop one. 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 predictive factors, usually refered to as potential predictors [7]. 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 **???**, or ones which could have a more general effect on overall health, such as socioeconomic and ethnicity variables **???**. 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 [19]. Prediction models should take into account as many predictive 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 [20].

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 **???**, [21], know as the Events-per-Variable (EPV). Recently, this number has been superseded by a methods to evaluate a specific required sample size [22]. If there aren't enough events to satisfy this criteria, then some potential predictors should be eliminated before any formal analysis takes place (for example using clinical knowledge) [23]. In general, it is also recommended that this development dataset contain at least 100 events (regardless of number of potential predictors) **???**, [15], [24]. A systematic review by Counsell et al [25] found that out of eighty-three prognostic models for acute stroke, less than 50% of them had more than 10 EPV, and the work by Riley et al [22] showed that less that [**Pull example from Riley 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 [9]. 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 [10], [26]. Derivation of the QRISK2 Score [27] (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 **???**, and its most modern incarnation, QRISK3, performed the external validation in the same paper [2] 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 [29], this validation was not performed at the same time as the initial development and is thus an external validation.

As with any technology, clinicians and researchers should be wary of models becoming outdated [30]. Healthcare systems and lifestyles change over time, and so models developed and externally validated in an outdated population will drift [31] and so should be updated regularly, as with QRISK [2] or automatically with a dynamic model [32]

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 [33]. 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 [3], [15]. Heterogeneity is particularly high when using multiple countries as a source of data [34] or if a potential predictor is of a subjective nature, which leads to discrepancies between assessors **???**. Overlooking of this clustering can lead to incorrect inferences [33]. 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 [11].

A prediction model researcher needs to select clinically relevant potential predictors for use in the development of the model [9]. 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 [9]. One key fact that many experts agree on is that categoriation of continuous predictors should be avoided [**Cite: LOADS**] as it retains much more predictive information. The cut-points of these categorisations lead to artificial jumps in the outcome risk [23]. It is also worth noting that cut-points are often either arbitrarily decided or data-driven with the latter leading to overfitting [23]. If categorisation is performed, clear rationale should be provided with an ackowledgement that this wil reduce performance **???**, [16]. When applying a model to a new population, extrapolation of a model should be avoided **???** 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 [20]. This ranges was later extended with the updated version **???** and currently can be applied to patients aged 25-84 [**Update with QRISK3**].

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 [23]. 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 [23]. 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 [9]. Univariables 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. Despite its clear weaknesses, any prognostic studies still use univariable analysis to build their models **???**.

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 untill all remaining variables are significant. Forward selection begins wth no variables and adds one a a time based on similar criteria. Under either of these methods, a lower significance level will exxlucde more variables [9]. Backward elimination of variables is preferable over forward selection as users are less likely to end up in local minima **???**. 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 [**Cite: p-values be bad reference**]. 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.

Model Validation

Impact Evaluation

1.2.4 Stratified Medicine

1.2.5 Examples

1.3 Competing Risks & Multi-State Models

1.4 Chronic Kidney Disease

1.4.1 Clinical Prediction Models

1.4.2 Multi-State Models

Chapter 2

The Application of Multi-State Methods to Develop Clinical Prediction Models Designed for Clinical Use - A Scoping Review

MA Barrowman, D Jenkins, GP Martin, N Peek, M Lambie, M Sperrin Last updated: 21 Apr

2.1 Introduction

eHealthcare is moving towards a more data-driven approach to decision making, exploiting the variety of data sources collected as part of routine care [1]. This increases efficiency, which is becoming increasingly vital as patients are living longer and requiring more care, while budgets are being reduced [2], [3]. Correspondingly, there has been a shift towards primary prevention, rather than purely treating disease as it arises [4] therefore clinical prediction models (CPMs) are more relevant than ever before [5].

Prognostic CPMs (those that predict the future) allow end-users to estimate an individual's probability/risk of experiencing an outcome of interest within a certain timeframe. CPMs are algorithms that relate a set of prognostic factors to the risk of a chosen outcome [6], often using multivariable regression. They can provide predictions of the future course of an illness and provide evidence for the commencement of medical interventions [7].

Along with this overall increase in importance, different methods of producing CPMs are also being used, and each makes different assumptions, and models at different levels of granularity. One of these methods is the Multi-State Model (MSM), an extension to traditional survival analysis wherein patients exist in one of many distinct states at any given time and can transition

between them (these individual transitions are akin to that of traditional survival analysis) [8]. A subset of MSMs is that of a Competing Risks model, where patients can only move from a single initial state to many absorbing states without any intermediate or transient states. A huge advantage of Multi-State CPMs, and indeed, Competing Risks CPMs, is that they can provide predictions for multiple outcomes with MSMs going further by allowing the prediction of multiple pathways to that outcome, whereas traditionally developed models only provide predictions for a single end-point.

However, little is known about how widely these types of models are implemented in clinically relevant prognostic research. Therefore, we here aim to document a scoping review protocol that will intend to uncover any prediction models using MSMs that have been developed for clinical use. As part of the process of this investigation, we will also document how many CPMs account for Competing Risks alone. We define a scoping review as described by Arksey and O'Malley [9] , which is similar to a systematic review, but with less formal outline for the analysis and synthesis of literature [10]. By assessing how MSMs have currently been applied in this field, we aim to describe the landscape of their current use, the context in which they are being used and discuss ways in which their use, application and uptake can be improved. To the best of our knowledge, a review such as this for Multi-State Models has never been performed.

2.2 Methods

2.2.1 Scope of Review

This review will cover articles related to the development of Multi-State Clinical Prediction models designed for clinical use. It will not include models that were developed solely for demonstrations of novel methodological improvements in the field of clinical prediction modelling and/or multi-state modelling. Article inclusion will be based on the screening of the article text and interpretation of its aims, primary distinction will be made on whether an existing dataset is used as a core part of the article or as a subsidiary example. It will include articles that validate previously developed models and those that review existing models, only so far as to use them to find the original development article (a method known as Snowballing).

As this analysis will follow the style of a scoping review; the final paper will adhere to the PRISMA-ScR guidelines [11], which were set out to extend the traditional PRISMA guidelines to a Scoping Review setting.

Models which focus only on a competing risks scenario (whether directly or simply adjusting for competing risks) will not be analysed in detail, however to avoid missing possible Multi-State Models, we will only omit these at the final stage of screening (See below). This will also allow for a brief description of how many CR models exist compared to the MSM models to be analysed in detail in this review.

As per the definitions set out by the PROGRESS research group, prognostic research is split into four overarching themes/types:

- Type I - Fundamental Prognosis Research [12]
- Type II - Prognostic Factor Research [13]
- Type III - Prognostic Model Research [14]
- Type IV - Stratified Medicine Research [15]

As such, we will be focusing on papers of Type III [14]. Articles related to the other types of prognostic research often develop a model within their work, but since the intent of these papers is to investigate overall outcomes, effects of an individual factor or interactive effects of treatments in individuals, they are considered disjoint from CPM development and so they will not be included in our analysis.

2.2.2 Initial Search Strategy

Search Terms

To ensure we cover as much of the medical literature as possible, we will use the Ovid search engine to search two databases:

- EMBASE (1974 to 2018 December 31)
- Ovid MEDLINE and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions 1946 to December 31, 2018

We will use a standard set of terms designed by Ingui & Rogers [16] and added to by Geersing et al [17] used for searching for clinical prediction related literature. We will also extend this by including search terms relating to time-to-event outcomes and/or survival analysis that were defined by the authors, and which aim to broaden our search (see table 1). This will be combined by a set of search terms designed to filter for MSMs and/or CRs.

These novel MSM/CR terms include “fine adj2 gray” to include papers which use the Fine & Gray subdistribution proportional hazard method [18]. It will also include“semimarkov or semi markov” to include articles which specify that the model adopts a semi-Markov perspective, which is common amongst MSMs [8]. However, we chose not to include the term “markov” alone as it is considered to be too unspecific to be of use (a la search for “model” alone when finding clinical prediction models). The full search details can be found in table 2.

We believe that the broadness of our search terms allows for high sensitivity in our results and will therefore provide a larger and more comprehensive pool of papers than using a more specific set of search terms.

[**Insert Table from paper**]

Validation set of articles

To ensure that our search strategy is satisfactory, we will compare our results to a set of Validation papers. These are papers that we are already aware of that satisfy our inclusion/exclusion criteria and which therefore should be included in our analysis. We will compare the results of our initial

search with this set of papers to ensure that all of the Validation set appear in our results. If they do not, then we will adjust our search strategy iteratively increasing sensitivity and improving the reach of our search until all Validation papers are included. The set of Validation papers is as follows:

- *Estimation and Prediction in a Multi-State Model for Breast Cancer*, Putter et al, 2006 [20]
- *A Multi-State Model to Predict Heart Failure Hospitalizations and All-Cause Mortality in Outpatients With Heart Failure With Reduced Ejection Fraction: Model Derivation and External Validation*, Upshaw et al, 2016 [21]
- *Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate*, Grams et al, 2018 [22]
- *Estimating transition probability of different states of type 2 diabetes and its associated factors using Markov model*, Nazari et al, 2018 [23]
- *Advantages of a multi-state approach in surgical research: how intermediate events and risk factor profile affect the prognosis of a patient with locally advanced rectal cancer*, Manzini et al, 2018 [24]

2.2.3 Filtering

Once the initial set of articles has been found, these will be filtered at various degrees of granularity to focus on papers which are included in the scope of our review as per our inclusion/exclusion criteria. We will also define which papers will be used only for the snowballing process, but will not be used as part of our analysis.

Inclusion/Exclusion Criteria

Inclusion

- Type III Prognostic Study Papers (i.e. those developing a clinical prediction model) [14]
- Papers which use a Multi-State Model framework to provide individual level patient predictions

Exclusion

- Papers that develop overall population level predictions (Type I)
- Papers focused on identification of prognostic factors (Type II)
- Papers that investigate stratified medicine (Type IV)
- Papers that only develop Competing Risks models
- Papers designed to describe methodological models with or without clinical application used only for an example

Stages

The filtering of the results will be performed in three stages: 1. Title (MB) 2. Abstract (MB with 20% replication by DJ) 3. Full Paper (MB with 20% replication by DJ)

Filtering will begin with an initial check through all titles to assess whether it is believed that the paper may be relevant to the review. This will help to omit a large amount of papers that were incorrectly returned by the broad search strategy. To ensure the review remains as sensitive as possible, only papers where it is abundantly clear that they violate an inclusion/exclusion criteria will be removed at this stage.

A second filter will be performed on the abstracts of the remaining articles and removed papers will be classified by the reason for their omission. To allow for faster data extraction, a final glancing filter will also be performed over the full papers to again reduce the numbers of collated papers in the final review and reduce the likelihood of removing papers at the analysis stage. To ensure robustness of this filtering, both of these stages will be replicated by a second reviewer (DJ) in a randomly selected 20% of the abstracts and papers and differences will be discussed internally. At this point, models focusing solely on competing risks (i.e. those without a transient state) will be filtered out.

2.2.4 Data Extraction

To study the use of Multi-State Clinical Prediction Models from a quantitative perspective, certain vital data points will be extracted from the extant models. These measurements can be grouped as to what element of the prediction model they are evaluating: * Clinically Relevant points * Number of patients * Clinical setting (i.e. primary vs secondary care, geographic setting) * Field of study (e.g. cardiovascular, renal, etc.) * Summary of patient demographics (i.e. inclusion/exclusion criteria) * Outcomes being predicted * Multi-State Model details * Number of States and what they are * Shape/Structure of the model (i.e. how patients can transition between states) * How were relevant variables chosen? * Transition assumptions (e.g. parametric vs non-parametric, PH assumption, etc...) * Stated justification for, and reported benefits of an MSM versus traditional methods. * Predictive Ability * Timeframe (e.g. single time point(s), continuous time prediction, dynamic prediction, etc...) * What validation was performed (None vs. Internal (bootstrap, CV, etc.) vs. External) * Comparisons to current guidelines * Assessment of Bias of their model (using PROBAST) * Utilisation of the TRIPOD Guidelines (e.g. Was it referenced? Was it adhered to?) * Prominence information * Number of citations (although not clinically relevant, it is relevant to understanding the model's utilisation) * Year of publication (again, not clinically relevant, but useful to spot any time trends in prominence and/or quality) The data extracted at this stage will be checked by DJ in 20% of the papers to confirm results for the analysis

2.2.5 Reporting

The search and filtering strategy will be depicted with a modified PRISMA flow diagram [30], which includes papers found by Snowballing and how they are included in the filtration process, see figure 1.

[**Add in PRISMA**]

A table of the extracted information will be included with the paper, depending on the number of results, this may be supplementary material. This information will also be summarised and analysed both quantitatively and qualitatively. For example, as the Illness-Death model [8] is simple and common amongst multi-state models, we will count how many of the MSCPMs use this structure as well as the other most common structures used. Any direct comparisons that can be made between predictions of this type (i.e. from the same field with the same outcomes) will be described.

Chapter 3

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 Last updated: 21 Apr

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 moreso 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 in Appendix A

3.1 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 [1]. 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 [2].

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 [3], 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 [4]. Therefore, if these variables are omitted form (or unavailable

for) the analysis of treatment effects in observational studies, then they can bias inferences [5]. 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 [6].

The issue of confounding in observational studies has been researched previously [7,8,9], 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 [10, 11]. Hence, it is widely recognised that sensitivity analyses are vital within the observational setting [12]. However these previous studies do not extend this work into a competing risk setting, meaning research in this space is lacking [13], 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 [14], 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 [15], 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.

3.2 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 [16].

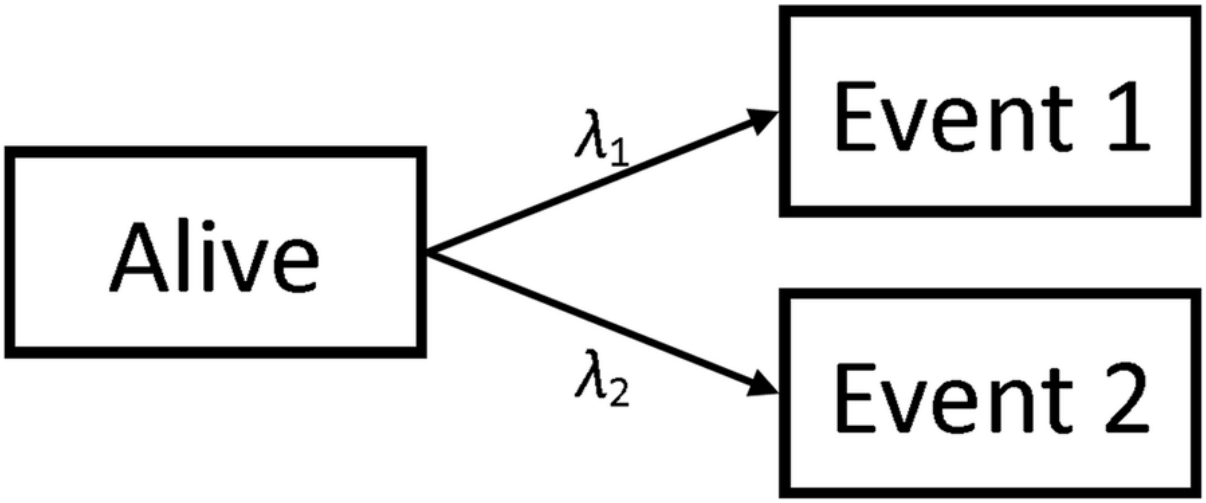
The data-generating mechanism defined two cause-specific hazard functions (one for each event), where the baseline hazard for event 1 was k times that of event 2, see Fig. @ref(fig:Transition_Diagram). We assumed a baseline hazard that was either constant (exponential distributed failure times), linearly increasing (Weibull distributed failure times) or biologically plausible [17]. The hazards used were thus:

$$\lambda_1(t|U,Z) = ke^{\beta_1 U+\gamma_1 Z}\lambda_0(t) \tag{3.1}$$

$$\lambda_2(t|U,Z) = ke^{\beta_2 U+\gamma_2 Z}\lambda_0(t) \tag{3.2}$$

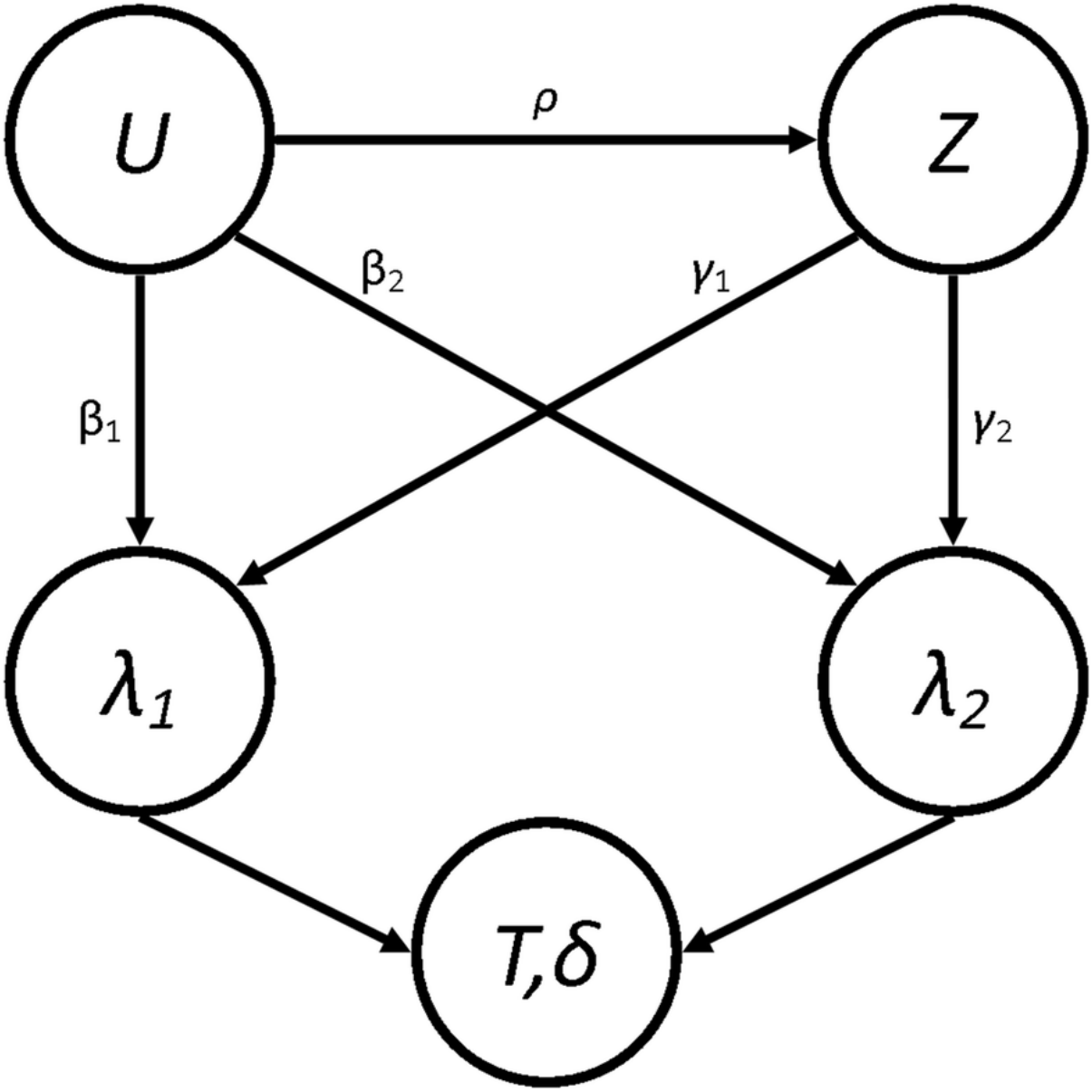
$$\lambda_0(t)\begin{cases} 1 & \text{Exponential} \\ 2t & \text{Weibull} \\ \exp -18 + 7.3t - 11.5t^{0.5}\log(t) + 9.5t^{0.5} & \text{Plausible} \end{cases} \tag{3.3}$$

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 [18].



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 $\Pr(Z = 1) = \pi$ [19]. Population end times and type of event were generated using the relevant hazard functions. The full process for the simulations can be found in Additional file 1. 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. @ref(fig:Model_DAG), 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 explicitly calculated what we would expect the true subdistribution treatment effects, Γ_1 and Γ_2, to be in these conditions [20] (See Additional file 2). 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 [18, 21] 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) [20].

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.
- $\theta_{\text{CSH},i} = \text{E}[\hat{\gamma}_i - \gamma_i]$ ~ 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 ($\gamma_2 = \gamma_2 = 0$) 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 ($\gamma_2 = \gamma_2 = -1$). 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 ($\gamma_2 = -1$) 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.

[**Insert Table 1**]

3.3 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 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).

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.

3.4 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 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 [22] 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, 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 [23] 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 [5] 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. [5] 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 [24]. They hypothesised that the association found elsewhere [25] 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. $\theta_{\mathrm{CSH},1} \neq 0$ 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. [26] 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. @ref(fig:Model_DAG), to visualise where unmeasured confounding may be a problem in analysis [27] and statisticians who deal with such models will use transition diagrams, see Fig. @ref(fig:Model_Transitions), to visualise potential patient pathways [28]. 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 [26] and Thompson [24]) 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 [14, 15, 29], 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. [30] in that censoring would provide less clarity of the presented results.

3.5 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.

Chapter 4

Inverse Probability Weighting Adjustment of the Logistic Regression Calibration-in-the-Large

MA Barrowman, A Pate, GP Martin, CJM Sammut-Powell, M Sperrin Last updated: 03 May

Abstract

Introduction

Methods

Results

Discussion

Supplementary Material

Supplementary Material is available in Appendix C.

4.1 Introduction

~~Clinical prediction models (CPMs)~~Clinical prediction models (CPMs) are statistical models/algorithms that aim to predict the presence (diagnostic) or furture occurence (prognostic) of an

event of interest, conditional on a set of predictor variables. Before they be implemented in practice, CPMs must be robustly validated. They need to be validated before they are used and a~~A~~ fundamental test of their validity is calibration: the agreement between observed and predicted outcomes. This requires that among individuals with $p\%$ risk of an event, $p\%$ of those have the event across the full risk range [35]. The simplest assessment of calibration is the calibration-in-the-large, which tests for agreement in mean calibration (the weakest form of calibration) [36]. 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 descibe how calibration changes across the risk range, such as calibration slope (see Appendix B). Calibration alone is not enough to fully assess a model's performance however and so we also need measures of discrimination (how well models discern between different patients), e.g the c-statistic and overall accuracy, e.g. the Brier Score.

~~In the case of Cox regression, however, estimation of calibration is complicated in three ways. First, calibration can be computed at multiple time-points and one must decide which time-points to evaluate, and how to integrate over these time-points. Second, there exists no explicit intercept in the model because of the non-parametric baseline hazard function [37]. Third, censoring needs to be handled in an appropriate way. 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 looked at integrated over time using martingale residuals [38]; however here we focus on the case where calibration at a specific time point is of interest - e.g. as is common in clinical decision support. The lack of intercept can be overcome provided sufficient information concerning the baseline survival curve is available (although this is rarely the case [39]. Once this is established, estimated survival probabilities are available. Censoring leads to problems in determining observed survival. This is commonly overcome by using Kaplan-Meier estimates [27], [37]. However the censoring assumptions required for the Kaplan-Meier estimate are stronger than those required for the Cox 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 [40] presents an alternative approach for calibration at external validation. He uses the approach of pseudo-observations, as described by Perme and Anderson [41] to overcome the censoring issue and produce observed probabilities at individual level; however, this assumes that censoring is independent of covariates. In this paper and another [42] he proposes the comparison of KM curves in 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.~~

In the case of time to event models, however, estimation of calibration is complicated in three ways. First, calibration can be computed at multiple time-points and one must decide which time-points to evaluate, and how to integrate over these time-points. 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 [38]; 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, there exists no explicit intercept in the model because of the non-parametric baseline hazard function [37]. The lack of intercept can be overcome provided sufficient information concerning the baseline survival curve is available (although this is rarely the case as seen in QRISK[**Cite:**], ASCVD[**Cite:**] and ASSIGN[**Cite:**]. Once this is established, estimated survival probabilities are available.

Third, censoring needs to be handled in an appropriate way. This is commonly overcome by using Kaplan-Meier estimates [27], [37], but the censoring assumptions required for the Kaplan-Meier estimate are stronger than those required for the Cox 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 [40], [42] 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 [27] with good results (6.34% predicted vs 6.25% observed in women and 8.86% predicted vs 8.88% observed in men), but ~~bad~~worse results in the QRISK3 update [2] (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) and an important change between the update was the lower age limit moved from 35 to 25, as well as the implementation of QRISK in clinical practice [**I remember discussing this with Alex & Matt a while ago as to whether the use of QRISK had a feedback loop when updated after it's own implementation. Did this go any further?**].

Royston [40] also presented an alternative approach for calibration at external validation. He uses the approach of pseudo-observations, as described by Perme and Anderson [41] 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 [43] 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 [44] who demonstrated that any proper scoring rule can be improved by the use of the IPCW. This work has been ~~added to~~extended by Han et al [45] and Liu et al [46] who demonstrated ~~that the c-statistic is also suitable~~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 first show, theoretically, how this method can be used and evidence that the metrics for calibration are amenable to its use. We then compare simulation results from using this weighted estimate to an unweighted estimate within various commonly used methods of calibration assessment.

4.2 Methods

4.2.1 Theory

[**Lots of Theory work on the probabilities. May need to drop this if we're unable to do it between us.**]

4.2.2 Aims

The aim of this simulation study is to ~~formalise~~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).

4.2.3 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 [47]. Each population was simulated with ~~two~~three parameters: β, γ and η, which defined the proportional hazards coefficients for the survival and censoring distributions and the baseline hazard function, respectively.

~~We varied the parameters to take all the values,$\gamma = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$, $\beta = \{-2, -1.5, -1, -0.5, 0.5, 1, 1.$ and $\eta = \{-{}^1/_2, 0, {}^1/_2\}$, that is the proportional hazard coefficients took the same values between -2 and 2, but β did not take the value of 0 because this would make a predictive model infeasible.~~

~~For each combination of parameters, we generated $N = 100$ populations of $n = 10,000$ patients (a high number of patients was chosen to avoid bias due to a small population sizeimprove precision of our estimates with~~Patients were generated with a single covariate $Z \sim N(0,1)$~~. For each patient,~~ 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 $e^{\gamma Z}$. 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 ~~the real-world~~practice, only Z, X and δ would be observed.

~~For each population, a prediction model for survival, F_P was chosen to be identical to the Data Generating Mechanism (DGM) to emulate a perfectly calibrated model (...)~~

~~This prediction model was used to generate an estimate of the Expected probability that a given patient, with covariate z, will have an event at the given time. To test the ability of approaches to detect miscalibration, 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 such (...)~~

~~The prediction models were assessed at 100 time points, evenly distributed between the 25th and 75th percentile of observed event times, X. At each time point, t, we removed patients who had been censored (i.e. $T < X_i$ & $\delta_i = 0$) and created an indicator variable for whether each patient had had the event yet or not: (...)~~

~~Similarly, we calculate a censoring prediction model, G, to be identical to the DGM: (...)~~

~~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:(...)~~

During each simulation, we varied the parameters to take all the values,$\gamma = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$, $\beta = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$ and $\eta = \{-{}^1/_2, 0, {}^1/_2\}$. 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)

4.2.4 Prediction Models

[**New section, taken from previous snippets, highlighting/strikethroughs will show the new changes**]

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

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

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

~~This~~These prediction models ~~was~~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.

4.2.5 The IPCW

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

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

4.2.6 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 [6], [40], [42], [48] 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 and this is compared with the average Expected value.
- Logistic Unweighted (LU) - Logistic regression is performed on the non-censored population to predict the binary Observed value using the logit(Expected) 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. ~~Logistic r~~Regression is performed ~~using~~with the complimentary log-log function as a link function and the log cumulative hazard as an offset ~~and~~with the Intercept ~~of the result is the estimater~~epresenting the estimate of calibration-in-the-large.

~~The weights within the LW method create a non-integer number of events within the regression. and the PO method can produce values that are not always 0 or 1 (as would be expected in an ordinary logistic regression). The values produced by PO will have to be artificially capped between 0 and 1, but otherwise these two methods do not cause any issues.~~Some of these methods produce unusual results for the regressions. Firstly, the weights within the LW method cause the "number of events" being processed (i.e the sum of the weighted events) to be non-integer. This is a minor issue and can be dealt with by most software packages [49]. Secondly, the PO method produces outcomes that are outside of the (0,1) range [41] required for the complimentary log-log function. To combat this, we re-scale the values produced to be with this range and perform the regression as normal.

4.2.7 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 $\theta_{i,U}(t)$.

4.2.8 Performance Measures

The measures we will take as performance measures as the Bias, the Empirical Standard Error ~~as~~and the Coverage at time, t, along with relevant standard errors and confidence intervals as per current recommendations [50]. These measures can be seen in table 4.1. For these estimates

Table 4.1: 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	$\hat{C}(t) = \frac{1}{N}\sum_{i=1}^{N} I\left(\theta_{i,L}(t) \leq \theta \leq \theta_{i,U}(t)\right)$	$\hat{C}_{SE}(t) = \frac{\hat{C}(t)\left(1-\hat{C}(t)\right)}{N}$

at each time point, Method and Model, the top and bottom 5% of all simulation estimates will be omitted, leaving $N = 90$ to avoid biasing the results from singly large random effects. The bias provides a measure 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.

~~For each estimand above, $\hat{Q}(t) = \{\hat{\theta}(t), \hat{E}(t), \hat{C}(t)\}$ and associated SE, $\hat{Q}_{\text{SE}}(t) = \{\hat{\theta}_{\text{SE}}(t), \hat{E}_{\text{SE}}(t), \hat{C}_{\text{SE}}(t)\}$, we average over time. As these measures will be taken at each of the 100 time points, $t_j : j = 1...100$, we summarise each of these measures as an average and as weighted average, as seen in table **??**. The weight used for the measure at time t_j is the average number of non-censored patients remaining in the population at time t_j, defined as n_j (note that this includes patients who have had the event).(. . .)~~

4.2.9 Software

All analysis was done in `R 3.6.3` [51] using the various `tidyverse` packages [52], Kaplan-Meier estimates were found using the `survival` package [53], Pseudo-Observations were evaluated with the `pseudo` package [54], and the results app was developed using `shiny`[55]. The code used for this simulation study is available on Github and the results can be seen in a shiny app

4.3 Results

[**Results shown here are new and improved from the previous version. No highlighting is shown**]

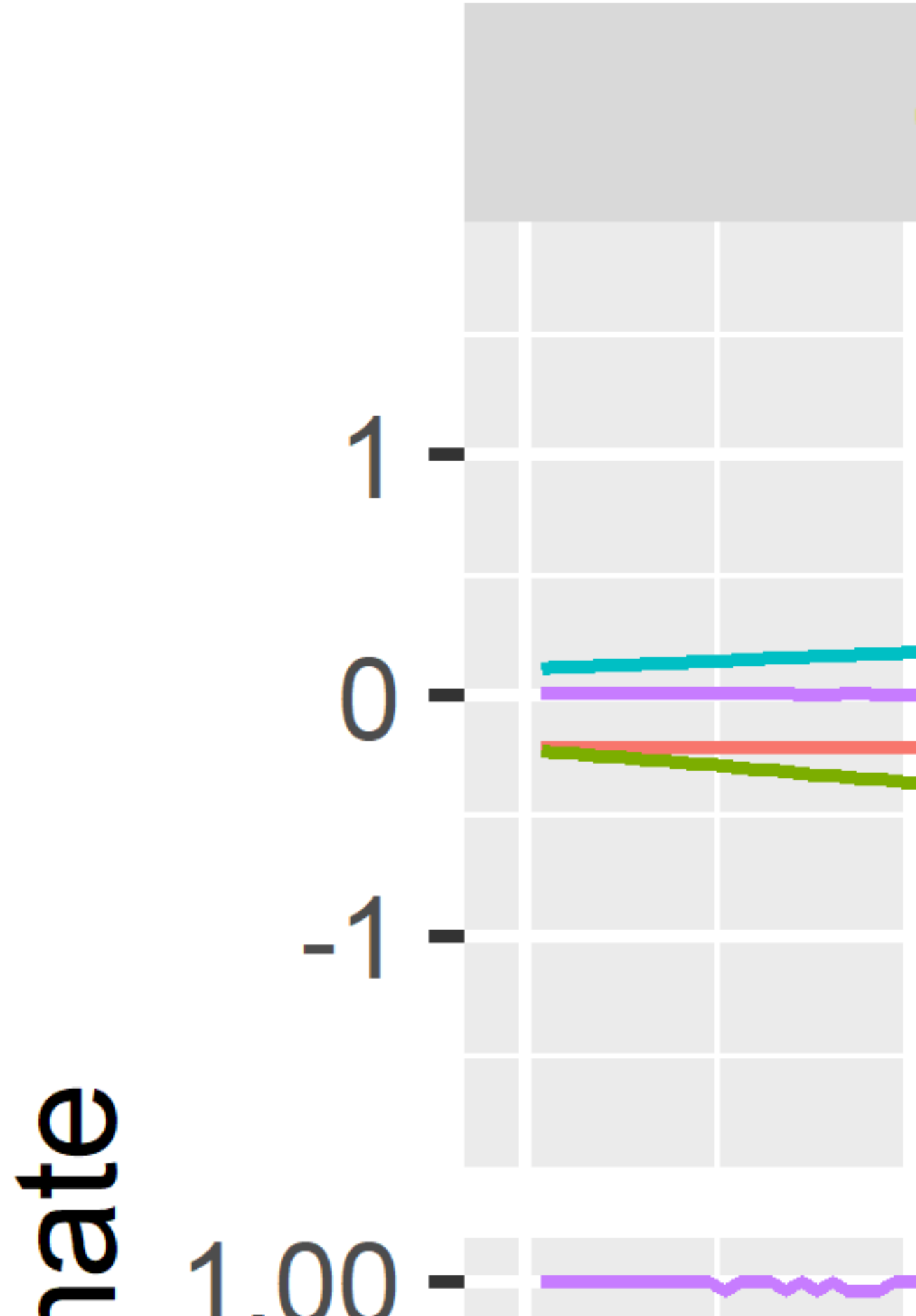

1
0
-1
1.00

Figure 4.1 shows the results when censoring is independent of covariates ($\gamma = 0$). The LW method provides strong coverage across the entire timeframe and miminal bias. The absolute bias for PO and LU increases over time with PO under-reporting the correct value and LO over-reporting. KM bias remains constant across the timeframe, but for the imperfect models, is constantly under- or over-reported. LU and PO also provide minimal coverage at all time points, whereas KM covers perfect in the early stages of the Perfect Model with coverage dropping off as time progresses. Empirical Standard Error is clse to 0 for all models.

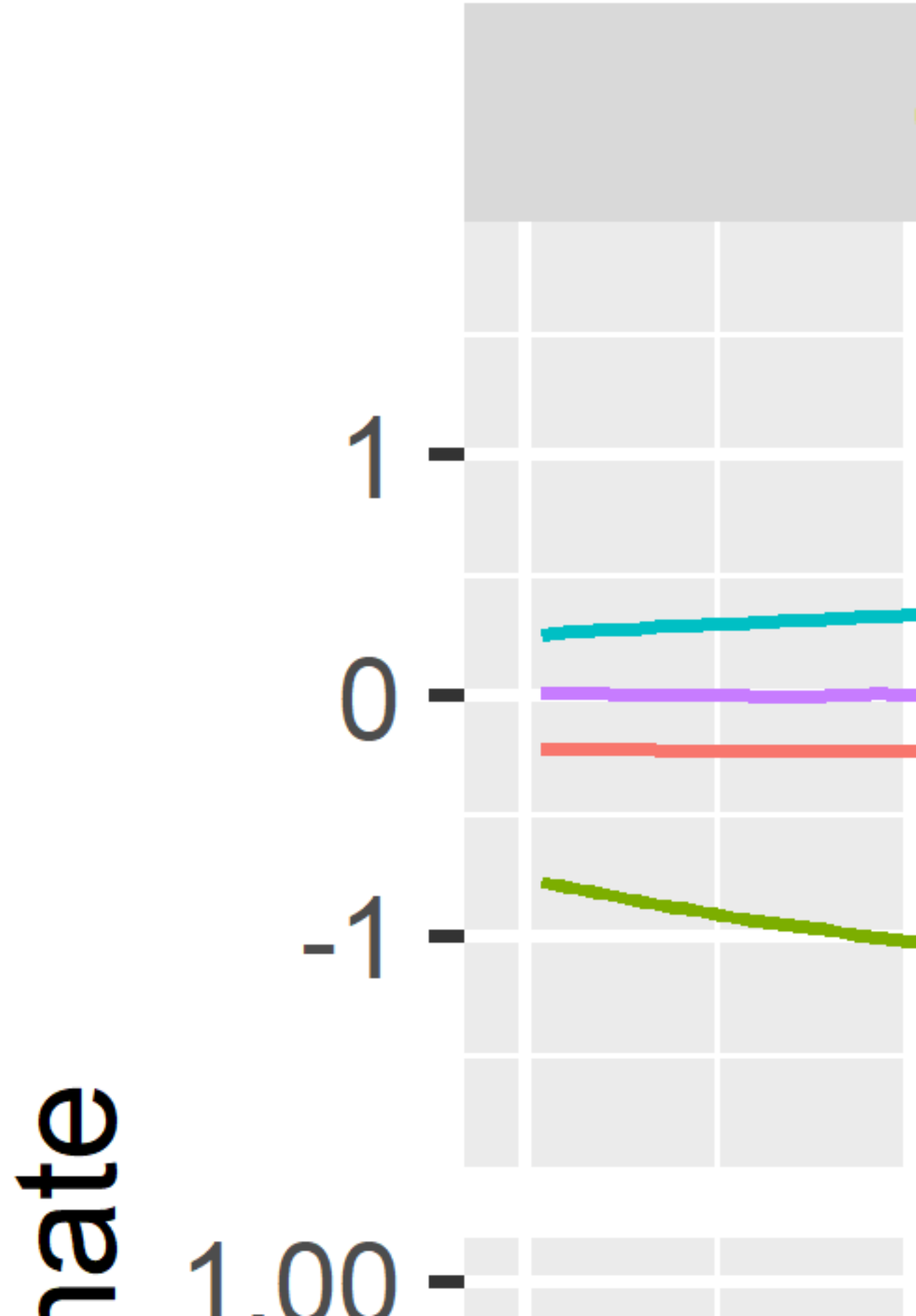

1
0
-1
1.00

Figure 4.2 shows the results when censoring and the event-of-interest have the same individual effects ($\beta = \gamma = 1$). The LW method provides strong coverage across the entire timeframe and miminal bias, although this coverage is reduced compared to the previous set of results shown (approximately 75% throughout). Once again, the absolute bias for PO and LU increases over time, however the under-reporting for PO is much more strongly pronounced. KM bias behaves similarly but for coverage, it starts off at around 50% coverage reaches a peak of full coverage approximately 25% of the way through the timeframe.

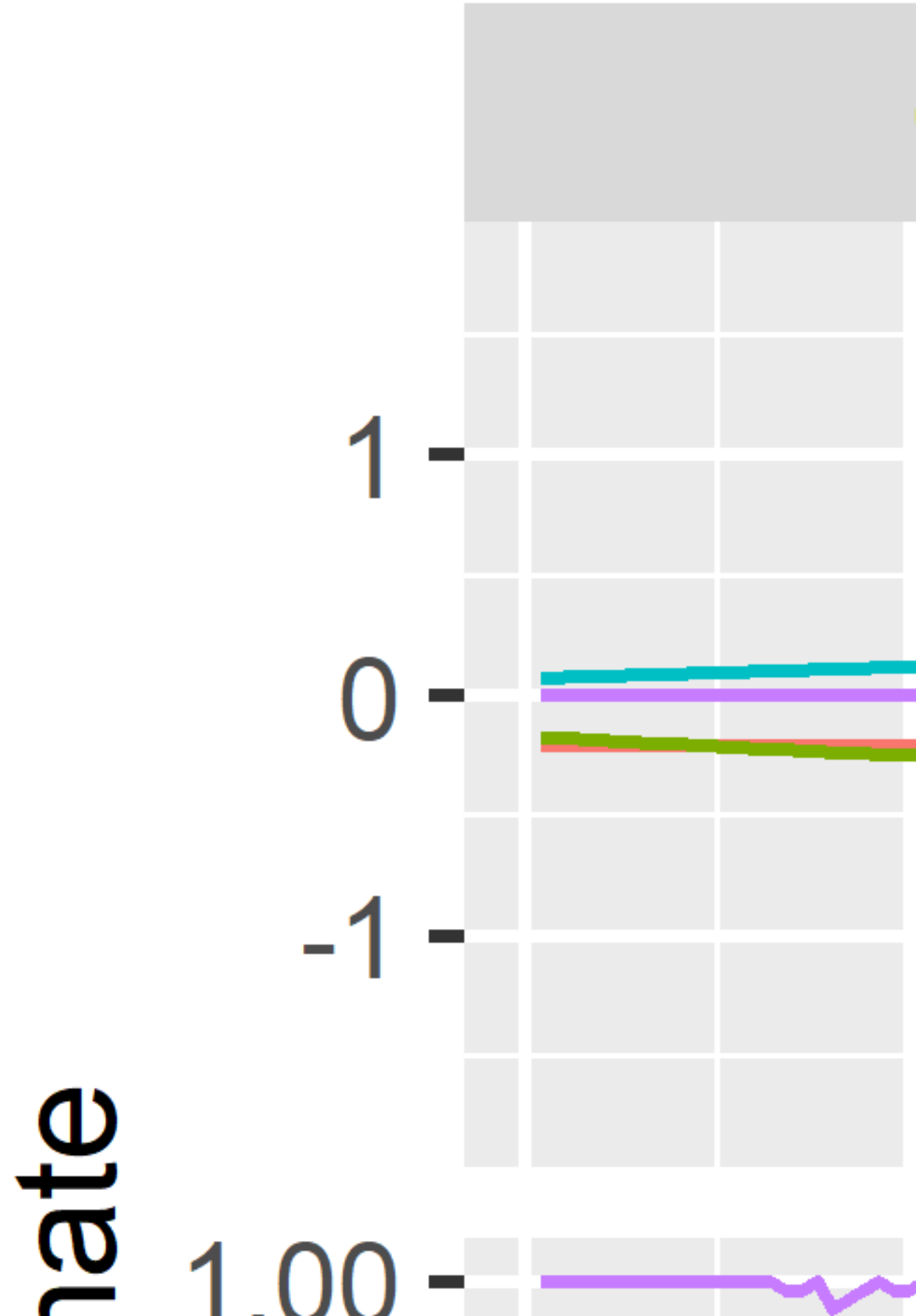

1
0
-1
1.00

Figure 4.3 shows the results when censoring and the event-of-interest have opposite individual effects ($\beta = 1, \gamma = -1$). The bias results are similar to those when censoring is independent. A difference here is that coverage begins greater than zero for the KM, LU and PO methods, but quickly drops to 0 before the 25% time point. For LW, the coverage appears to reduce to around 80% by the end of the time point.

4.4 Discussion

Weighting = Good.

Not Weighting = Bad.

limitation: Maybe the "True" θ for the under and over predictions were wrong and that would explain the low Coverage.

Chapter 5

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

MA Barrowman, GP Martin, N Peek, M Lambie, M Sperrin Last updated: 30 Apr

5.1 Introduction

Clinical Prediction Models (CPMs) provide individualised risk of a patient's outcome (cite), based on that patient's predictors. These predictions will usually be in the form of a risk score or probability. However, using traditional modelling techniques, these CPMs will only predict a single outcome. Multi-State Clinical Prediction Models (MS-CPMs) combine the multi-state modelling framework to the prognostic field to provide predictions for multiple outcomes in a single model. Once a CPM has been developed, it is important to assess how well the model actually performs (cite). This process is called Model Validation and involves comparing the predictions produced by the model to the actual outcomes experienced by patients (cite). 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 (cite). Models can also be validated on a novel dataset (external validation), which is used to assess the generalisability and transportability of the model (cite). 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 (cite) or to analyse how

well a model discriminates between patients, we could use the c-statistic (cite). The current metrics that are commonly used have been designed and extended to work in a variety of model development frameworks. However, these extensions are limited to either a single outcome (as in traditionally developed models) or do not adequately account for the censoring of patients (as commonly occurs in longitudinal 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 are collapsible to the current metrics and that they adjust for the bias induced by 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 [cite something]. Other methodologists have extended existing metrics to multinomial outcomes [cite van Calster], which do not contain a time-based component; to simple competing risks scenarios [cite CR c-statistic], which do not contain transient states; or to [... insert third relevant example]. Spitoni et al [cite Spitoni 2018]] 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. It is the hope of the authors that this work will increase the uptake of multi-state models and the sub-field of MS-CPMs will grow appropriately.

5.2 Motivating Data Set

[**Table One for The Glasgow Data**]

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 [cite Dev/Valid Paper]. The model was developed using data from the Salford Kidney Study (SKS) and then applied to an external dataset derived from the West of Scotland (see Table 2) [1]. The original model predicts the probability that a patient has begun RRT and/or died after their first recorded eGFR below 60 ml/min/1.73m2, 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. The Three-State model used in our example is designed as an Illness-Death Model [2], 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 1).

[**Figure of the MSM**]

[**Describe Glasgow Data**]

5.3 Current Approaches

Here we describe three commonly used performance metrics for assessing the performance of a traditional survival clinical prediction model. 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 longitudinal 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) [3]. This allows the uncensored patient population to be representative of the entire patient population.

5.3.1 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 noninformative 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 and 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 a Non-Informative 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, which many models can also be compared to as models as close to this display excellent predictive abilities. Although models may perform worse than a non-informative one, we will not consider these in detail here as they are considered to be without worth in terms of predictive ability. 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. In order to allow for simplicity and understanding of these measures, they will be standardised to the same scales.

5.3.2 Notation

Throughout this paper, we will use consistent notation which is shown here for reference and to avoid repetition in definitions, etc...

[**Notation Table**]

5.3.3 Patient Weighting

[Lots of formula, so will leave for now]

5.3.4 Accuracy - Brier Score

5.3.5 Discrimination - c-statistic

5.3.6 Calibration - Intercept and Slope

5.4 Extension to Multi-State Models

5.4.1 Trivial Extensions

5.4.2 Accuracy - Multiple Outcome Brier Score

5.4.3 Discrimination - Polytomous Discriminatory Index

Computational Limitations

5.4.4 Calibration - Multinomial Intercept, Matched and Unmatched Slopes

5.5 Application to Real-World Data

5.5.1 Accuracy

5.5.2 Discrimination

5.5.3 Calibration

5.6 Discussion

Chapter 6

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 Last updated: 29 Apr

Abstract

Introduction

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 multiple outcomes with a single model can be advantageous. Multi-State Models are a method to provide these kinds of predictions.

Methods

We developed a Multi-State Clinical Prediction Model (MSCPM) using tertiary care data from the Salford Kidney Study as our development data set and secondary care data from the West of Scotland (SERPR) dataset as our validation set. We developed three models of different levels of complexity; a Two-State Model (Alive and Dead), a Three-State Model (Untreated CKD, Renal Replacement Therapy and Dead) and a Five-State model (Untreated CKD, Haemodialysis, Peritoneal Dialysis, Transplant and Dead). We used Royston-Parmer regression techniques to allow us to provide individualised predictions for patients. Model performance was assessed for accuracy, discrimination and calibration using methods both internally and externally. The best performing model was used to produce a CPM Calculator for clinical use.

Results

Of the three models produced, Age was a strong predictor of mortality in all cases and outcomes were highly dependent on primary renal diagnosis. Models performed well in both the internal and external validation with the Three-State Model out performing overall. The Three-State Model was used to develop the online Calculator.

Discusssion

Our CPMs provide clinicians and patients with multiple outcome predictions. This implies that users of these models can get more information about their potential future without a loss to the quality of that prediction.

Supplementary Material

Supplementary Material is available in Appendix C.

6.1 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 condition, more specifically, a prognostic model allows the prediction of future events [8]. CPMs use data from previous patients to estimate the outcomes of an individual patient. Prognostic models are used in clinical practice to influence treatment decisions such as the prescribing of statins for cardiovascular disease via the application of the QRISK models [2].

Within Chronic Kidney Disease (CKD), prognostic models have been developed to predict mortality [56]–[60], ESRD [57], the commencements of RRT [59], [61]–[63] or mortality after beginning dialysis [64]–[66]. Some previous models have used the commencement of RRT as a proxy for ESRD [67]–[69], while others have investigated the occurrence of cardiovascular events

within CKD patients[70]–[72]. Reviews by Grams & Coresh [73], Tangri et al [74] and Ramspek et al [75], 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], [76].

Methodologically, the majority of existing CKD prediction models fail to account for completing events [58], [60], [77], have high risks of bias [56], [57], [61] or are otherwise flawed compared to modern clinical prediction standards [8], [13]

In 2013, Begun et al [78] 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 [79] 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 [63] developed an MSM focusing on the categories of Calculated Panel Reactive Antibodies (CPRA) and kidney transplant and/or death.

Most recently, in 2018, Grams et al [80] 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 [17], [81]–[98]. These kinds of cut-offs can be useful when informing patients and clinicians of a patient's diagnosis and to coincide with policy, but inherently cause a loss of information when done before the data analysis stage and so these models go against current statistical recommendations [17], [81]–[98]. These kinds of assumptions are also subject to measurement error [99] and interval censoring [100], i.e. we do not know when exactly when a patient moved from CKD Stage III to CKD Stage IV, or whether drop in estimated Glomerular Function Rate (eGFR) was temporary or inaccurate. For example, Kulkarni [63] assumes that a patient with an CPRA of (5%) is the same as a patient with an CPRA of (75%) and that a patient with an CPRA of (89.9%) is vastly different from a patient with an CPRA of (90%). Moreover, none of these models have undergone any validation process, whether internal or external [10].

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 [9].

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 a pair of multinomial logistic models analysed at 2 and 4 years. The major downside of this model is that it can only produce predictions at those predefined time points and it assumes homogeneity of transition

times. For example, the first model assumes that a patient who began RRT 1 month after study entry is the same as one who began after 1 year & 11 months into the study and then the second model assumes these patients are the same as one who begins RRT at 3 years and 11 months.

Therefore, the aim of this study was to improve on previous efforts to model a patient's pathways through a Multi-State Model by choosing transition points which can be exactly identified and include states which produce a drastic difference in patient characteristics. Our modeling techniques allow for individual predictions (using a proportional hazards model) of multiple outcomes (using MSMs) at any time point (using cubic splines). The models produced by this process will then be validated, both internally and externally, to compare their results and demonstrate the transportability of the (statistically robust) clinical prediction models. We report our work in line with the TRIPOD guidelines for development and validation of clinical prediction models [13], [14].

6.2 Methods

6.2.1 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 [101], [102] 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 [103]) 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 (SIR) [104]. 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 at any time during their CKD journey.

This allows for a wider range of baseline eGFR measurements and patients who have been suffering from CKD, translating to a model which can be applied to

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. 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 available within SIR and so 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 [105] 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 [80]. 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
- Patient is not enduring an AKI [106], [107].

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.

6.2.2 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 6.1). 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 before and after the transition (i.e. Tx was modelled as a second absorbing state).

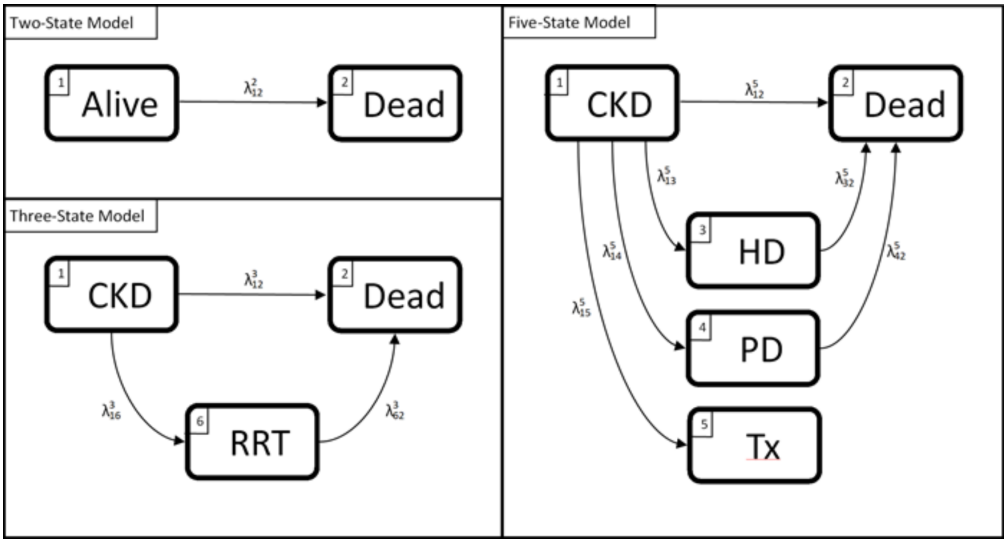


Figure 6.1: Diagram of the three models, the states being modelled and relevant transitions

The models were developed and validated as discussed in the Supplementary materials in appendix C.

6.2.3 Example

Once the models have been developed, we will apply them to two example patients to demonstrate their use and applicability to the general population. We will provide a direct clinical 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 6.1. Out 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.

Table 6.1: Details of the Example Patients

	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	Chronic Obstructive Pulmonary Disease Liver Disease Solid Tumour		Diabetes Chronic Obstructive Pulmonary Disease Hypertension

6.2.4 Calculator

As part of this work, we also intend to produce 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` [51] using the various `tidyverse` packages [52], as well as the `mice` [108], `flexsurv` [109], `nnet` [110] and `furrr` [111] packages. The calculator was produced using the `shiny` package [55].

6.3 Results

6.3.1 Data Sources

As seen in table 6.2, The Age of both populations were centred around 64-65 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) . Blood pressure was also consistent across populations (140/75 vs 148/76 for development vs validation). The blood test results (Corrected Calcium, Albumin, Haemoglobin and Phosphate) was close together, with the further difference being Haemoglobin with an average of 123 in SKS and 109 in SERPR and a much larger standard deviation in SERPR compared to SKS (38 vs 17). 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. Similar to the eGFR measures, the uPCR results were similar, but the rates of change were much broader in the validation dataset compared to the SKS and were generally increasing, whereas SKS remained stationary (73 vs 0). Levels of missingness were much higher in the SERPR dataset in most continuous variables.

Table 6.2: Population demographics for the continuous variables presented as: mean (IQR) [min,max] <number missing (percent missing)>

	SKS (Development)	SERPR (Validation)
Age		
Age	64.378 (19.000) [20.000, 94.000] < 0 (0.00%)>	65.880 (17.000) [18.000, 98.000] < 0 (0.00%)>
eGFR		
eGFR[a]	30.368 (22.386) [3.577, 59.965] < 0 (0.00%)>	36.132 (21.444) [1.651, 59.998] < 0 (0.00%)>
eGFR Rate[a]	-0.015 (0.293) [-19.107, 33.781] <1,278 (42.87%)>	-25.476 (44.229) [-8,755.272, 9,260.375] < 0 (0.00%)>
uPCR		
uPCR[f]	0.112 (0.103) [0.000, 2.025] < 245 (8.21%)>	0.184 (0.147) [0.000, 6.390] <7,513 (96.76%)>
uPCR Rate[f]	-0.096 (0.188) [-70.727, 28.198] <1,777 (59.61%)>	73.177 (0.384) [-2.255, 3,051.403] <7,721 (99.44%)>
Measures		
SBP[b]	140.193 (29.000) [77.000, 220.000] < 50 (1.67%)>	147.746 (33.000) [82.000, 258.000] <6,880 (88.61%)>
DBP[b]	74.555 (14.000) [36.000, 159.000] < 52 (1.74%)>	76.263 (18.000) [35.000, 128.000] <6,879 (88.60%)>
BMI[c]	28.848 (7.842) [13.182, 61.466] < 572 (19.18%)>	29.331 (7.851) [15.343, 48.301] <7,681 (98.93%)>
Albumin[d]	42.152 (5.000) [12.000, 52.000] < 60 (2.01%)>	36.490 (6.000) [7.000, 53.000] <3,455 (44.50%)>
Corrected Calcium[e]	2.302 (0.180) [1.209, 3.660] < 68 (2.28%)>	2.408 (0.160) [1.419, 3.610] <5,113 (65.85%)>
Haemoglobin[d]	122.977 (23.000) [61.000, 195.000] < 72 (2.41%)>	108.588 (30.000) [6.250, 208.000] <3,968 (51.10%)>
Phosphate[e]	1.162 (0.320) [0.430, 3.710] < 87 (2.91%)>	1.203 (0.320) [0.370, 4.370] <5,127 (66.03%)>

[a] (ml/min/1.73m^2) or per year [b] (mmHG) [c] (kg/m^2) [d] (g/l) [e] (mmol/l) [f] (g/mmol) or per year

Table 6.3 shows a breakdown of the categorical variables across the populations. In the development population, there are far more males than females, whereas in the validation population the proportions are much more matched. 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. The majority of the SKS patients were former smokers, however this information was unavailable in the SERPR dataset. Primary Renal Diagnosis suffered from very high levels of missingness in the validation dataset, but was much better recorded in the development dataset (although still far from perfect). Overall, there were high levels of comorbidities within the SKS

Table 6.3: Population demographics for the categorical variables presented as number (percent)

	SKS (Development)	SERPR (Validation)
Gender		
Male	1,865 (62.56 %)	3,915 (50.42 %)
Female	1,116 (37.43 %)	3,849 (49.57 %)
Ethnicity		
White	2,875 (96.44 %)	683 (8.79 %)
Asian	75 (2.51 %)	12 (0.15 %)
Black	21 (0.70 %)	7 (0.09 %)
Other	10 (0.33 %)	2 (0.02 %)
<Ethnicity Missing>	0 (0.00 %)	7,060 (90.93 %)
Smoking Status		
Former	1,535 (51.49 %)	
Non-Smoker	979 (32.84 %)	
Smoker	379 (12.71 %)	
Former 3Y	46 (1.54 %)	
<Smoking Status Missing>	42 (1.40 %)	
Primary Renal Diagnosis		
Systemic diseases affecting the kidney	1,304 (43.74 %)	299 (3.85 %)
Glomerular disease	442 (14.82 %)	225 (2.89 %)
Tubulointerstitial disease	268 (8.99 %)	164 (2.11 %)
Miscellaneous renal disorders	227 (7.61 %)	188 (2.42 %)
Familial / hereditary nephropathies	173 (5.80 %)	102 (1.31 %)
<Renal Diagnosis Missing>	567 (19.02 %)	6,786 (87.40 %)

population as shown in table 6.4, but these levels were much lower in the SERPR population, possibly due to the data extraction processed (where data is un-recorded, no history is assumed). In SKS, most comorbidities were at over 80% prevalence, apart from diabetes mellitus, which had a lower prevalence of 33% and over 97% (2,891) patients had a history of liver disease. In SERPR, hypertension was the highest prevalence in SERPR at 40% (3,122), followed by diabetes mellitus at 20% (1,546) and cerebrovascular accident was the lowest prevalence at 2.36% (184). Liver disease, chronic obstructive pulmonary disease and solid tumour data were unavailable in the SERPR data. 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 followup time in SKS was 15.0 years and in SERPR it was 10.1 years. This information can be seen in table 6.5.

Table 6.4: Population comorbidity prevalence for the two populations presented as number (percent) <number missing (percent missing)>

	SKS (Development)	SERPR (Validation)
Diabetes (DM)	992 (33.32%) < 4 (0.13%)>	1,546 (19.91%) < 0 (0.00%)>
Congestive Cardiac Failure (CCF)	2,414 (81.08%) < 4 (0.13%)>	406 (5.22%) < 0 (0.00%)>
Prior Myocardial Infarction (MI)	2,492 (83.70%) < 4 (0.13%)>	556 (7.16%) < 0 (0.00%)>
Ischemic Heart Disease (IHD)	2,393 (80.38%) < 4 (0.13%)>	867 (11.16%) < 0 (0.00%)>
Peripheral Vascular Disease (PVD)	2,485 (83.47%) < 4 (0.13%)>	376 (4.84%) < 0 (0.00%)>
Prior Cerebrovascular Accident (CVA)	2,727 (91.60%) < 4 (0.13%)>	184 (2.36%) < 0 (0.00%)>
Chronic Obstructive Pulmonary Disease (COPD)	2,411 (80.98%) < 4 (0.13%)>	
Chronic Liver Disease (LD)	2,891 (97.11%) < 4 (0.13%)>	
Solid Tumour (ST)	2,570 (86.32%) < 4 (0.13%)>	
Hypertension (HT)	2,546 (91.48%) <198 (6.64%)>	3,122 (40.21%) < 0 (0.00%)>

Table 6.5: Event times for the two populations presented as Number of Events Median (Inter-Quartile Range) [Max]

Transition	SKS (Development)		SERPR (Validation)	
Two				
Alive to Dead	1,427	3.9 y (4.3 y) [15.0 y]	3,025	4.9 y (3.3 y) [10.1 y]
Three				
CKD to Dead	1,125	3.5 y (4.2 y) [15.0 y]	2,579	4.8 y (3.2 y) [10.1 y]
CKD to RRT	680	2.5 y (3.3 y) [14.1 y]	1,130	3.8 y (3.8 y) [10.1 y]
RRT to Dead	302	2.2 y (3.2 y) [13.5 y]	446	1.5 y (2.4 y) [9.1 y]
CKD to Dead	1,125	3.5 y (4.2 y) [15.0 y]	2,579	4.8 y (3.2 y) [10.1 y]
CKD to HD	344	2.5 y (3.5 y) [14.1 y]	887	3.8 y (3.7 y) [10.1 y]
CKD to PD	229	2.0 y (2.9 y) [12.9 y]	149	3.5 y (4.1 y) [9.6 y]
CKD to Tx	107	3.2 y (2.7 y) [12.1 y]	94	4.8 y (4.5 y) [9.7 y]
HD to Dead	185	2.0 y (3.2 y) [11.8 y]	398	1.5 y (2.5 y) [9.1 y]
PD to Dead	107	2.3 y (3.2 y) [11.7 y]	47	2.1 y (2.3 y) [8.5 y]

6.3.2 Example

The example patients seen in Table 6.1 were passed through our Three-State prediction model and the results for all time-points are shown in figure 6.2. 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, 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.

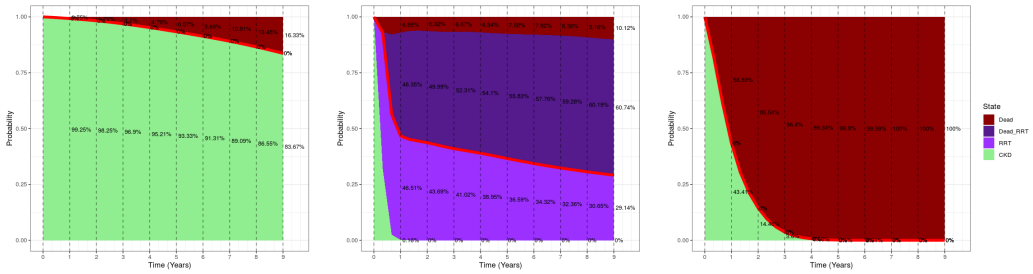


Figure 6.2: Results of Example Patients

6.3.3 Calculator

The calculator is available online here:

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

6.4 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 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.

Models performed well in model validation with the Three-State Model slightly out performing the other two models in calibration and overall predictive ability, however the Five-State model performed marginally better in terms of discriminative ability. Both Multi-State Models outperformed the Two-State (Traditional) Model.

The application of a Multi-state clinical prediction model to this field is novel and gives a powerful tool for providing individualised predictions of multiple outcomes at a wide range of time points. 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. The Renal Department at SRFT is a tertiary care facility for CKD 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.

In the Five-State Model, We omitted the analysis of the Tx to Dead state due to the anticipated low number of events within the SKS dataset. The lowest number of events for a transition was therefore PD to Dead, which had only 107. Altogether, we considered 26 covariates (with 4 categorical covariates) and so this equates to 36 predictor parameters and an events per predictor parameter (EPP) of 2.97. This is below the recommendations of Riley et al [22], whose calculations produce a requirement of 4.54 EPP. This requirement was also not satisfied by the CKD to PD transition (EPP = 6.36,required = 10.2) or the CKD to Tx transition (EPP = 2.97, required = 17.6). Fortunately, this limitation is confined to the Five-State Model.

We have assumed a proportional hazards relationship between the predictors and probability of survival, which is considered by some to be a strong assumption to make, however we acknowledge this limitation, and the authors believe that it is mitigated by the flexibility that the assumption permits. In addition to the general PH assumption, the R-P model requires the assumption that the log cumulative hazard function follows a cubic spline, (however this is a much weaker assumption [112]), which is modelled as part of the regression. We did not assess the viability of these models as it was believed this assumption to make our results more understandable.

Compared to the raw internal validation, the model performance during the external validation was worse for all metrics. However, once adjusted for optimism, the results were much more cohesive which implies that the model is highly 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.

Although not directly assessing causality in regards to state-transitions, our Three-State model can be used by clinicians to either expedite or delay transition of a patient onto RRT, if it is believed that this would be beneficial. Alternatively, the Five-State Model can be interpreted to provide information regarding *which* treatment might be benficial for a patient.

Our paper has clearly demonstrated the accuracy of such a model. However, further research would be needed to establish the effectiveness and efficacy of its use in clinical practice [113] 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 well in terms of accuracy, calibration and discrimination when applied internally and externally. This shows directly that the models are suitable for use in populations similar to both our development and our validation datasets. It can also be concluded that the models can be transported and applied to any population with a

similar healthcare system to the UK.

Chapter 7

Conclusion

Last updated: 21 Apr

Here is where my concluding section will go.

The end.

Appendix A

How unmeasured confounding in a competing risks setting can affect treatment effect estimates in observational studies - Supplementary Material

A.1 Simulation Details

The populations

A.2 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 proportio

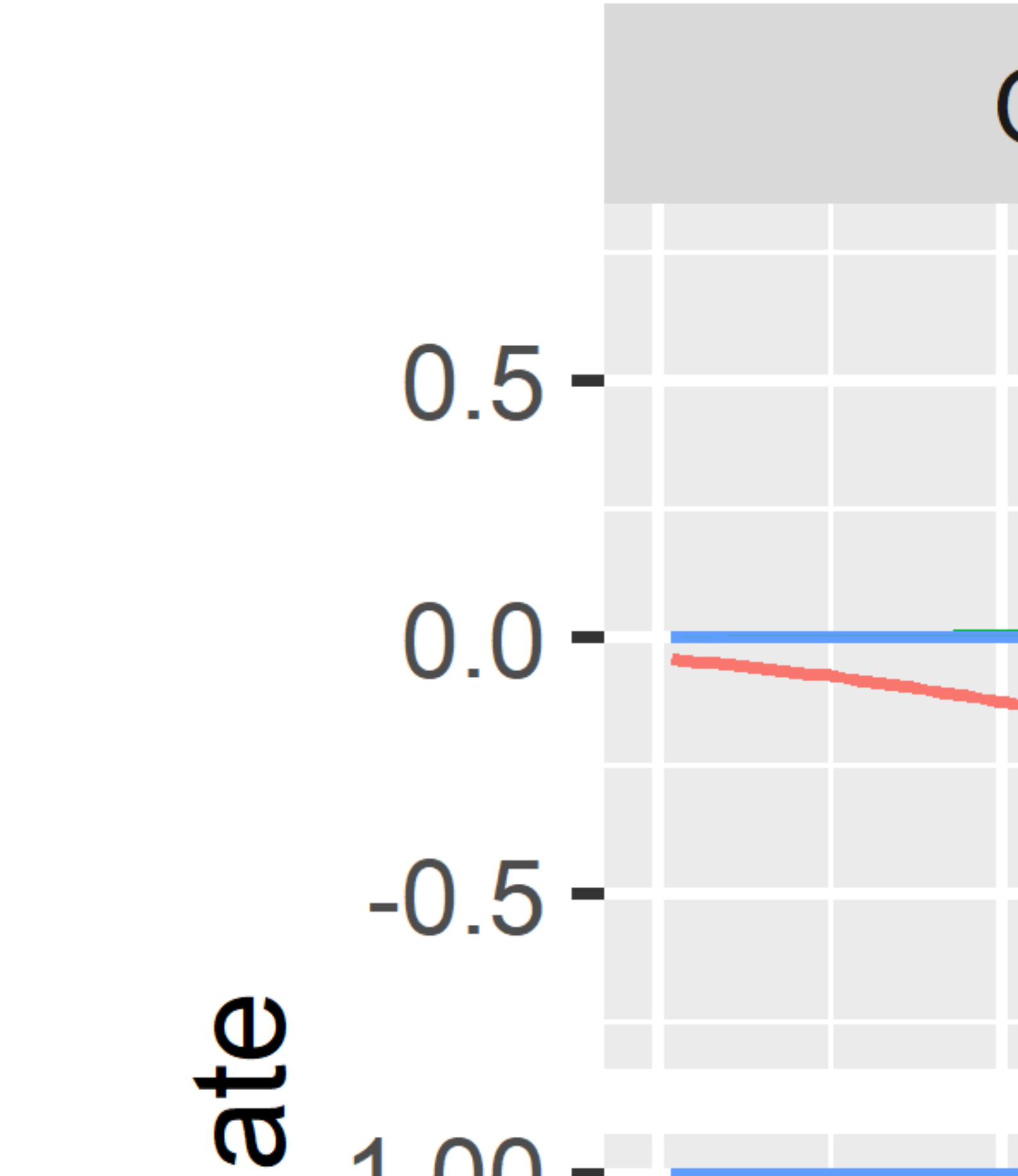
Appendix B

Inverse Probability Weighting Adjustment of the Logistic Regression Calibration-in-the-Large - Supplementary Material

B.1 Calibration Slope

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. The Logistic Weighted, Logistic Unweighted and Pseudo-Observation methods described above 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).

B.1.1 Results



Results currently show bias/coverage/EmpsE away from 0, rather than 1. Needs fixing. Oops.

B.1.2 Discussion

Brief discussion, much briefer than the main points.

Appendix C

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

C.1 Statistical Analysis

C.1.1 Development

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). 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 [114]). Ethnicity was assessed in the populations, but as most patients were white, it was omitted as a potential predictor from the models.

uPCR and eGFR Rate of change were also calculated [115], [116] as the difference between the two most recent measures divided by time difference in years. Age^2, log(Age), log(eGFR Rate) and log(uPCR Rate) were considered as transformations within the model. log(Calendar Time) was included as a covariate to adjust for secular trends in treatment preferences [31]. Calendar Time was defined as length of time between start date and 1st January 2019.

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 [117]. Each transition was modelled under a proportional hazards assumption using the Royston-Parmar technique [112] to estimate coefficients for each covariate and a restricted cubic spline (on the log-time scale) for the baseline cumulative hazard. The cumulative hazards for each transition can be combined to produce estimates for the probability of a patient being in any state at any time [118].

For variable selection, we stacked the imputed datasets together to create a larger, pseudo-population [119] and performed backwards-forwards selection based on minimising the AIC at each step. This was repeated for each transition and for different numbers of evenly spaced knots in modelling the form of the baseline hazard, K={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 [120]. This gave a model fully defined by smooth cubic splines representing the cumulative cause-specific hazard and individualised proportional hazards for each transition.

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 [121]. 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 [122].

C.1.2 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[123]. The bootstrapping method was also used for both validations 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 $< 30/\text{ml/min}/1.73\text{m}^2$.

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 **??**(chap-performance-metrics)

The overall accuracy of each model was assessed using the MSM adjusted Brier Score, which is a proper score function [124] 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 [125]. 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) [126] which extends the logistic regression to three or more mutually exclusive outcomes [6]. 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.

C.2 Model Results

C.2.1 Two State Model

Table C.1 shows the proportional hazard ratios for the transitions in the Two-State Model. Older patients have a higher hazard towards death, low adn decreasing eGFR increased hazard as did a history of diabetes. Patients with a primary renal diagnosis included in the ERA-EDTA [114] definition of Systemic diseases affecting the kidney had the highest likelihood of death. Equation (C.1) below shows the baseline cumulative hazard functions for the transition from Alive to Dead

Table C.1: Proportional Hazards for each transition in the Two-State Model

Age
(Age-60)
(Age-60)
log(Age)
eGFR
eGFR
eGFR Rate
log(eGFR Rate)
uPCR
uPCR
uPCR Rate
log(uPCR Rate)
Measures
SBP
DBP
BMI
Albumin
Corrected Calcium
Haemoglobin
Phosphate
Gender
Female
Smoking Status
Former (3 years+)
Non-Smoker
Smoker
Primary Renal Diagnosis
Familial / hereditary nephropathies
Glomerular disease
Miscellaneous renal disorders
Tubulointerstitial disease
Comorbidity
DM
CCF
MI
IHD
PVD
CVA
COPD
LD
ST
HT

in the Two-State Model.

$$\Lambda_{0,16}(t) = \begin{cases} 1.15764\log(t) + 6.17808 & 0 \leq t < 3 \\ 0.00075\log(t)^3 - 0.00247\log(t)^2 + 1.16036\log(t) + 6.17708 & 3 \leq t < 520 \\ 0.13983\log(t)^3 - 2.61165\log(t)^2 + 17.47602\log(t) - 27.83122 & 520 \leq t < 984 \\ -0.19361\log(t)^3 + 4.28227\log(t)^2 - 30.03426\log(t) + 81.30981 & 984 \leq t < 1454 \\ -0.15458\log(t)^3 + 3.42966\log(t)^2 - 23.82553\log(t) + 66.23901 & 1454 \leq t < 2009 \\ 0.37097\log(t)^3 - 8.56171\log(t)^2 + 67.37553\log(t) - 164.97265 & 2009 \leq t < 2900 \\ -0.16209\log(t)^3 + 4.18769\log(t)^2 - 34.26862\log(t) + 105.14551 & 2900 \leq t < 5497 \\ 1.79561\log(t) + 1.61764)) & 5497 \leq t \end{cases} \tag{C.1}$$

Table C.2 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

Table C.2: Internal Validation of the Two-State Model, results presented as Estimate (95% CI, where possible)

Predicting	eGFR	One Year	Two Year	Five Year	Average
Brier					
Two	< 60	0.63 (0.62, 0.63)	0.69 (0.69, 0.69)	0.66 (0.66, 0.67)	0.63 (0.62, 0.63)
Two	< 30	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					
Two	< 60	0.82 (0.82, 0.82)	0.85 (0.84, 0.85)	0.81 (0.81, 0.81)	0.81 (0.81, 0.82)
Two	< 30	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					
Two	< 60	0.01 (0.00, 0.01)	0.01 (0.00, 0.01)	-0.02 (-0.02, -0.01)	-0.00 (-0.01, -0.00)
Two	< 30	-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					
Two	< 60	*1.33*	*1.46*	*1.26*	*1.48*
Two	< 30	*1.23*	*1.25*	*1.30*	*1.51*

C.3 shows the results from the external validation in the Two-State Model, which shows similar results to the internal validation with slightly impaired perfomance, which is to be expected in an external validation.

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

Predicting	eGFR	One Year	Two Year	Five Year	Average
Brier					
Two	< 60	0.64 (0.63, 0.64)	0.57 (0.56, 0.57)	0.57 (0.56, 0.58)	0.56 (0.56, 0.57)
Two	< 30	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					
Two	< 60	0.81 (0.81, 0.82)	0.81 (0.80, 0.81)	0.80 (0.79, 0.80)	0.78 (0.78, 0.78)
Two	< 30	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					
Two	< 60	-0.00 (-0.00, 0.00)	0.02 (0.01, 0.02)	0.00 (0.00, 0.01)	-0.00 (-0.00, 0.00)
Two	< 30	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					
Two	< 60	*1.29*	*1.25*	*1.72*	*2.21*
Two	< 30	*1.37*	*1.37*	*2.05*	*1.88*

C.2.2 Three State Model

In the Three-State Model, 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. The full results are shown in table C.4.

Table C.4: Proportional Hazards for each transition in the Three-State Model

	CKD to Dea
Age	
(Age-60)	0.161 (-0.051, 0.374
(Age-60)	-0.000 (-0.002, 0.000
log(Age)	-5.725 (-17.969, 6.518
eGFR	
eGFR	-0.013 (-0.019, -0.006
eGFR Rate	
log(eGFR Rate)	0.042 (-0.125, 0.210
uPCR	
uPCR	0.125 (-0.318, 0.569
uPCR Rate	
log(uPCR Rate)	
Measures	
SBP	-0.001 (-0.004, 0.002
DBP	0.006 (0.000, 0.013
BMI	
Albumin	-0.044 (-0.064, -0.024
Corrected Calcium	0.280 (-0.192, 0.752
Haemoglobin	-0.013 (-0.017, -0.008
Phosphate	0.511 (0.132, 0.890
Gender	
Female	-0.235 (-0.371, -0.099
Smoking Status	
Former (3 years+)	-0.212 (-0.879, 0.453
Non-Smoker	-0.198 (-0.345, -0.051
Smoker	0.356 (0.160, 0.551
Primary Renal Diagnosis	
Familial / hereditary nephropathies	-0.424 (-0.854, 0.006
Glomerular disease	-0.394 (-0.635, -0.154
Miscellaneous renal disorders	-0.263 (-0.505, -0.021
Tubulointerstitial disease	-0.463 (-0.741, -0.184
Comorbidity	
DM	0.122 (-0.011, 0.255
CCF	-0.394 (-0.535, -0.253
MI	-0.246 (-0.397, -0.094
IHD	0.102 (-0.041, 0.245
PVD	-0.248 (-0.394, -0.103
CVA	-0.070 (-0.252, 0.111
COPD	-0.289 (-0.433, -0.145
LD	-0.169 (-0.578, 0.239
ST	-0.274 (-0.431, -0.117
HT	

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 etiology were strongly associated with the transitions giving a wide range of predictions.

The equations (C.2), (C.3) and (C.4) shows the baseline cumulative hazard functions for the transition from CKD to Dead, CKD to RRT and RRT to Dead, respectively in the Three-State Model.

$$\Lambda_{0,16}(t) = \begin{cases} 1.19795\log(t) + 17.68798 & 0 \leq t < 3 \\ -9e-05\log(t)^3 + 3e-04\log(t)^2 + 1.19761\log(t) + 17.68811 & 3 \leq t < 443 \\ 0.15869\log(t)^3 - 2.9019\log(t)^2 + 18.88018\log(t) - 18.22403 & 443 \leq t < 873 \\ -0.30096\log(t)^3 + 6.43659\log(t)^2 - 44.36299\log(t) + 124.54338 & 873 \leq t < 1295 \\ -0.04158\log(t)^3 + 0.86028\log(t)^2 - 4.40166\log(t) + 29.08554 & 1295 \leq t < 1876 \\ 0.51263\log(t)^3 - 11.67048\log(t)^2 + 90.03919\log(t) - 208.17245 & 1876 \leq t < 2738 \\ -0.23992\log(t)^3 + 6.19863\log(t)^2 - 51.3924\log(t) + 164.96467 & 2738 \leq t < 5497 \\ 1.98997\log(t) + 11.72244)) & 5497 \leq t \end{cases} \tag{C.2}$$

$$\Lambda_{0,15}(t) = \begin{cases} 1.55753\log(t) - 5.44635 & 0 \leq t < 18 \\ -0.00279\log(t)^3 + 0.0242\log(t)^2 + 1.48757\log(t) - 5.37894 & 18 \leq t < 270 \\ -0.13576\log(t)^3 + 2.25776\log(t)^2 - 11.01817\log(t) + 17.96111 & 270 \leq t < 538 \\ 0.49133\log(t)^3 - 9.57146\log(t)^2 + 63.36209\log(t) - 137.93609 & 538 \leq t < 919 \\ -0.76978\log(t)^3 + 16.24541\log(t)^2 - 112.80782\log(t) + 262.78176 & 919 \leq t < 1316 \\ 0.30039\log(t)^3 - 6.81452\log(t)^2 + 52.82254\log(t) - 133.77073 & 1316 \leq t < 2000 \\ -0.0123\log(t)^3 + 0.31552\log(t)^2 - 1.37092\log(t) + 3.53243 & 2000 \leq t < 5173 \\ 1.32717\log(t) - 4.15818)) & 5173 \leq t \end{cases} \tag{C.3}$$

$$\Lambda_{0,56}(t) = \begin{cases} 1.35522\log(t) - 7.7618 & 0 \leq t < 8 \\ -0.01704\log(t)^3 + 0.1063\log(t)^2 + 1.13417\log(t) - 7.60859 & 8 \leq t < 196 \\ 0.21761\log(t)^3 - 3.6103\log(t)^2 + 20.75671\log(t) - 42.14228 & 196 \leq t < 506 \\ -0.67558\log(t)^3 + 13.07415\log(t)^2 - 83.12956\log(t) + 173.47479 & 506 \leq t < 816 \\ 0.8043\log(t)^3 - 16.69103\log(t)^2 + 116.42771\log(t) - 272.49495 & 816 \leq t < 1388 \\ -1.26732\log(t)^3 + 28.27738\log(t)^2 - 208.94656\log(t) + 512.26643 & 1388 \leq t < 1927 \\ 0.17019\log(t)^3 - 4.34244\log(t)^2 + 37.78861\log(t) - 109.83234 & 1927 \leq t < 4940 \\ 0.85568\log(t) - 5.12598)) & 4940 \leq t \end{cases} \tag{C.4}$$

Table C.5 shows the results from the internal validation in the Three-State Model. Performance was overall slightly better in patients in the <60 eGFR group than in the <30 eGFR group. All measures degraded over time, but the average scores remained strong. Table C.6 shows the

Table C.5: Internal Validation of the Three-State Model, results presented as Estimate (95% CI, where possible)

Predicting	eGFR	One Year	Two Year	Five Year	Average
Brier					
Three	< 60	0.74 (0.74, 0.75)	0.68 (0.68, 0.69)	0.64 (0.64, 0.65)	0.67 (0.67, 0.68)
Three	< 30	0.75 (0.74, 0.75)	0.73 (0.73, 0.73)	0.68 (0.67, 0.68)	0.68 (0.67, 0.68)
Two	< 60	0.75 (0.75, 0.75)	0.75 (0.75, 0.76)	0.67 (0.67, 0.67)	0.67 (0.67, 0.68)
Two	< 30	0.71 (0.71, 0.72)	0.72 (0.72, 0.73)	0.65 (0.65, 0.66)	0.67 (0.67, 0.68)
c-statistic					
Three	< 60	0.87 (0.87, 0.87)	0.84 (0.84, 0.85)	0.84 (0.84, 0.84)	0.83 (0.83, 0.84)
Three	< 30	0.87 (0.86, 0.87)	0.84 (0.84, 0.84)	0.84 (0.84, 0.84)	0.83 (0.83, 0.84)
Two	< 60	0.86 (0.86, 0.86)	0.86 (0.86, 0.86)	0.83 (0.83, 0.84)	0.83 (0.83, 0.84)
Two	< 30	0.86 (0.85, 0.86)	0.86 (0.85, 0.86)	0.85 (0.85, 0.85)	0.84 (0.83, 0.84)
Intercept					
		-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)	-0.02 (-0.02, -0.01)	0.00 (0.00, 0.00)
Three	< 60	0.00 (0.00, 0.01)	-0.01 (-0.02, -0.01)	-0.01 (-0.02, -0.01)	0.00 (-0.00, 0.00)
		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	-0.00 (-0.01, -0.00)	0.00 (-0.00, 0.00)
		-0.01 (-0.01, -0.01)	0.00 (-0.00, 0.00)	-0.02 (-0.02, -0.01)	0.00 (-0.00, 0.00)
Three	< 30	-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, 0.00)	0.03 (0.02, 0.03)	-0.00 (-0.01, -0.00)	0.00 (-0.00, 0.00)
Two	< 60	-0.00 (-0.01, -0.00)	0.02 (0.01, 0.02)	0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)
Two	< 30	-0.00 (-0.00, 0.00)	-0.04 (-0.04, -0.04)	0.00 (0.00, 0.01)	-0.00 (-0.00, 0.00)
Slope					
		1.25, 0.00, 0.04	*1.17*, -0.06, 0.01	*1.21*, -0.02, -0.01	*1.32*, -0.01, 0.00
Three	< 60	-0.03, *1.10*, -0.00	0.03, *1.25*, -0.01	0.01, *1.44*, -0.04	-0.00, *1.37*, -0.01
		-0.00, 0.01, *1.16*	-0.01, 0.01, *1.37*	-0.02, 0.02, *1.27*	-0.00, 0.00, *1.33*
		1.21, 0.02, 0.07	*1.36*, -0.00, -0.01	*1.35*, -0.04, -0.01	*1.31*, 0.00, 0.00
Three	< 30	-0.04, *1.24*, 0.07	0.00, *1.31*, 0.03	0.04, *1.33*, 0.04	-0.01, *1.33*, 0.01
		0.01, 0.01, *1.16*	-0.02, -0.00, *1.26*	0.04, -0.02, *1.34*	-0.00, 0.00, *1.35*
Two	< 60	*1.21*	*1.28*	*1.27*	*1.31*
Two	< 30	*1.05*	*1.21*	*1.21*	*1.34*

results from the external validation in the Three-State Model.

Table C.6: External Validation of the Three-State Model, results presented as Estimate (95% CI, where possible)

Predicting	eGFR	One Year	Two Year	Five Year	Average
Brier					
Three	< 60	0.69 (0.68, 0.69)	0.70 (0.69, 0.70)	0.61 (0.60, 0.61)	0.62 (0.62, 0.63)
Three	< 30	0.68 (0.67, 0.68)	0.72 (0.71, 0.72)	0.65 (0.64, 0.65)	0.63 (0.62, 0.63)
Two	< 60	0.67 (0.67, 0.67)	0.70 (0.69, 0.70)	0.63 (0.63, 0.63)	0.62 (0.62, 0.63)
Two	< 30	0.66 (0.66, 0.67)	0.70 (0.70, 0.70)	0.65 (0.65, 0.66)	0.63 (0.62, 0.63)
c-statistic					
Three	< 60	0.82 (0.82, 0.82)	0.83 (0.83, 0.83)	0.79 (0.79, 0.79)	0.81 (0.80, 0.81)
Three	< 30	0.85 (0.84, 0.85)	0.84 (0.84, 0.84)	0.83 (0.83, 0.83)	0.81 (0.81, 0.81)
Two	< 60	0.85 (0.85, 0.86)	0.84 (0.84, 0.85)	0.80 (0.80, 0.80)	0.81 (0.80, 0.81)
Two	< 30	0.83 (0.83, 0.83)	0.82 (0.82, 0.82)	0.80 (0.80, 0.81)	0.81 (0.81, 0.81)
Intercept					
		0.00 (-0.00, 0.00)	0.01 (0.01, 0.02)	0.05 (0.04, 0.05)	-0.00 (-0.00, 0.00)
Three	< 60	0.00 (0.00, 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)	0.00 (0.00, 0.01)	0.00 (0.00, 0.00)	0.00 (0.00, 0.01)
		0.04 (0.04, 0.05)	0.02 (0.01, 0.02)	0.01 (0.01, 0.01)	0.00 (-0.00, 0.00)
Three	< 30	0.01 (0.00, 0.01)	-0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
		0.00 (0.00, 0.01)	0.01 (0.01, 0.02)	0.00 (0.00, 0.01)	-0.00 (-0.00, 0.00)
Two	< 60	0.00 (-0.00, 0.00)	-0.05 (-0.05, -0.04)	0.01 (0.01, 0.02)	-0.00 (-0.00, 0.00)
Two	< 30	0.02 (0.01, 0.02)	-0.00 (-0.00, 0.00)	0.01 (0.00, 0.01)	-0.00 (-0.00, 0.00)
Slope					
		1.35, -0.04, 0.04	*1.13*, -0.00, 0.08	*1.45*, -0.02, 0.03	*1.54*, 0.00, -0.00
Three	< 60	0.06, *1.49*, -0.03	-0.02, *1.31*, 0.01	0.03, *1.73*, 0.03	0.01, *1.52*, 0.00
		0.01, 0.00, *1.25*	-0.03, 0.01, *1.39*	0.04, -0.00, *1.47*	-0.00, 0.00, *1.54*
		1.23, 0.03, 0.00	*0.99*, -0.01, -0.00	*1.56*, 0.00, 0.00	*1.62*, -0.01, 0.01
Three	< 30	-0.05, *1.20*, -0.04	0.04, *1.34*, -0.03	-0.05, *1.49*, 0.03	-0.00, *1.53*, 0.01
		-0.00, -0.00, *1.37*	0.00, 0.01, *1.46*	0.03, 0.01, *1.63*	0.00, -0.00, *1.58*
Two	< 60	*1.28*	*1.26*	*1.64*	*1.51*
Two	< 30	*1.18*	*1.25*	*1.57*	*1.59*

C.2.3 Five State Model

Table C.7 shows the proportional hazard ratios for the transitions in the Five-State Model.

Table C.7: Proportional Hazards for each transition in the Five-State Model

	CKD to Dea
Age	
(Age-60)	0.161 (-0.051, 0.374
(Age-60)	-0.000 (-0.002, 0.000
log(Age)	-5.725 (-17.969, 6.518
eGFR	
eGFR	-0.013 (-0.019, -0.006
eGFR Rate	
log(eGFR Rate)	0.042 (-0.125, 0.210
uPCR	
uPCR	0.125 (-0.318, 0.569
uPCR Rate	
log(uPCR Rate)	
Measures	
SBP	-0.001 (-0.004, 0.002
DBP	0.006 (0.000, 0.013
BMI	
Albumin	-0.044 (-0.064, -0.024
Corrected Calcium	0.280 (-0.192, 0.752
Haemoglobin	-0.013 (-0.017, -0.008
Phosphate	0.511 (0.132, 0.890
Gender	
Female	-0.235 (-0.371, -0.099
Smoking Status	
Former (3 years+)	-0.212 (-0.879, 0.453
Non-Smoker	-0.198 (-0.345, -0.051
Smoker	0.356 (0.160, 0.551
Primary Renal Diagnosis	
Familial / hereditary nephropathies	-0.424 (-0.854, 0.006
Glomerular disease	-0.394 (-0.635, -0.154
Miscellaneous renal disorders	-0.263 (-0.505, -0.021
Tubulointerstitial disease	-0.463 (-0.741, -0.184
Comorbidity	
DM	0.122 (-0.011, 0.255
CCF	-0.394 (-0.535, -0.253
MI	-0.246 (-0.397, -0.094
IHD	0.102 (-0.041, 0.245
PVD	-0.248 (-0.394, -0.103
CVA	-0.070 (-0.252, 0.111
COPD	-0.289 (-0.433, -0.145
LD	-0.169 (-0.578, 0.239
ST	-0.274 (-0.431, -0.117
HT	

PPE DIX . EVALUATION OF MENINGEAN VALVE DISEASE MODELS - SUPP

The equations (C.6), (C.7), (C.8) and (C.5) show the baseline cumulative hazard functions from the CKD state to HD, PD, Tx and Dead, respectively. Equation (C.9) shows the baseline cumulative hazard function from HD to Dead and Equation (C.10) shows the baseline cumulative hazard function from PD to Dead.

$$\Lambda_{0,16}(t) = \begin{cases} 1.19795\log(t) + 17.68798 & 0 \leq t < 3 \\ -9e-05\log(t)^3 + 3e-04\log(t)^2 + 1.19761\log(t) + 17.68811 & 3 \leq t < 443 \\ 0.15869\log(t)^3 - 2.9019\log(t)^2 + 18.88018\log(t) - 18.22403 & 443 \leq t < 873 \\ -0.30096\log(t)^3 + 6.43659\log(t)^2 - 44.36299\log(t) + 124.54338 & 873 \leq t < 1295 \\ -0.04158\log(t)^3 + 0.86028\log(t)^2 - 4.40166\log(t) + 29.08554 & 1295 \leq t < 1876 \\ 0.51263\log(t)^3 - 11.67048\log(t)^2 + 90.03919\log(t) - 208.17245 & 1876 \leq t < 2738 \\ -0.23992\log(t)^3 + 6.19863\log(t)^2 - 51.3924\log(t) + 164.96467 & 2738 \leq t < 5497 \\ 1.98997\log(t) + 11.72244)) & 5497 \leq t \end{cases} \tag{C.5}$$

$$\Lambda_{0,12}(t) = \begin{cases} 2.10248\log(t) - 8.40415 & 0 \leq t < 22 \\ -0.02708\log(t)^3 + 0.25114\log(t)^2 + 1.3262\log(t) - 7.60432 & 22 \leq t < 398 \\ 0.1382\log(t)^3 - 2.71684\log(t)^2 + 19.09196\log(t) - 43.05188 & 398 \leq t < 939 \\ -0.08093\log(t)^3 + 1.78288\log(t)^2 - 11.70775\log(t) + 27.2208 & 939 \leq t < 1680 \\ 0.00603\log(t)^3 - 0.15477\log(t)^2 + 2.68263\log(t) - 8.40353 & 1680 \leq t < 5173 \\ 1.35914\log(t) - 4.63106)) & 5173 \leq t \end{cases} \tag{C.6}$$

$$\Lambda_{0,13}(t) = \begin{cases} 1.54717\log(t) - 2.95572 & 0 \leq t < 18 \\ -0.0099\log(t)^3 + 0.08586\log(t)^2 + 1.29901\log(t) - 2.71663 & 18 \leq t < 399 \\ 0.06206\log(t)^3 - 1.20706\log(t)^2 + 9.04226\log(t) - 18.17463 & 399 \leq t < 1143 \\ -0.02442\log(t)^3 + 0.61979\log(t)^2 - 3.82137\log(t) + 12.01805 & 1143 \leq t < 4720 \\ 1.42179\log(t) - 2.7669)) & 4720 \leq t \end{cases} \tag{C.7}$$

$$\Lambda_{0,14}(t) = \begin{cases} 1.2532\log(t) - 26.76925 & 0 \leq t < 57 \\ 0.05585\log(t)^3 - 0.67736\log(t)^2 + 3.99179\log(t) - 30.46001 & 57 \leq t < 673 \\ -1.53678\log(t)^3 + 30.432\log(t)^2 - 198.5656\log(t) + 409.16648 & 673 \leq t < 968 \\ 1.6017\log(t)^3 - 34.30313\log(t)^2 + 246.5168\log(t) - 610.87894 & 968 \leq t < 1408 \\ -1.63794\log(t)^3 + 36.16075\log(t)^2 - 264.35866\log(t) + 623.7665 & 1408 \leq t < 1874 \\ 0.33702\log(t)^3 - 8.48952\log(t)^2 + 72.12758\log(t) - 221.49141 & 1874 \leq t < 4432 \\ 0.84439\log(t) - 21.97914)) & 4432 \leq t \end{cases} \tag{C.8}$$

$$\Lambda_{0,26}(t) = \begin{cases} 1.29244\log(t) - 8.98979 & 0 \leq t < 8 \\ -0.01397\log(t)^3 + 0.08716\log(t)^2 + 1.1112\log(t) - 8.86417 & 8 \leq t < 204 \\ 0.26151\log(t)^3 - 4.30937\log(t)^2 + 24.49956\log(t) - 50.33752 & 204 \leq t < 479 \\ -0.93294\log(t)^3 + 17.80365\log(t)^2 - 111.9599\log(t) + 230.35972 & 479 \leq t < 764 \\ 1.30576\log(t)^3 - 26.7818\log(t)^2 + 184.02359\log(t) - 424.6091 & 764 \leq t < 1217 \\ -1.24908\log(t)^3 + 27.66815\log(t)^2 - 202.79666\log(t) + 491.39983 & 1217 \leq t < 1828 \\ 0.18579\log(t)^3 - 4.66437\log(t)^2 + 40.05781\log(t) - 116.63908 & 1828 \leq t < 4310 \\ 1.02311\log(t) - 7.74927)) & 4310 \leq t \end{cases} \tag{C.9}$$

$$\Lambda_{0,36}(t) = \begin{cases} 1.55188\log(t) - 8.40785 & 0 \leq t < 9 \\ -0.02867\log(t)^3 + 0.18901\log(t)^2 + 1.13657\log(t) - 8.10367 & 9 \leq t < 167 \\ 0.24473\log(t)^3 - 4.00711\log(t)^2 + 22.60308\log(t) - 44.70978 & 167 \leq t < 553 \\ -0.99362\log(t)^3 + 19.45322\log(t)^2 - 125.54725\log(t) + 267.14313 & 553 \leq t < 875 \\ 1.22972\log(t)^3 - 25.73085\log(t)^2 + 180.53981\log(t) - 424.02428 & 875 \leq t < 1410 \\ -1.86859\log(t)^3 + 41.67205\log(t)^2 - 308.23775\log(t) + 757.44556 & 1410 \leq t < 1937 \\ 0.31596\log(t)^3 - 7.93077\log(t)^2 + 67.19201\log(t) - 189.72839 & 1937 \leq t < 4302 \\ 0.83657\log(t) - 4.66673)) & 4302 \leq t \end{cases} \tag{C.10}$$

Table C.8 shows the results from the internal validation in the Five-State Model. The calibration slope results are shown in a seperate table for both the internal and external validation. Table C.9 shows the results from the external validation in the Five-State Model. Table C.10 shows the calibration slopes for the model in the internal and external datasets in both the < 60 eGFR and <30 eGFR sub-populations.

Table C.8: Internal Validation of the Five-State Model, results presented as Estimate (95% CI, where possible)

Predicting	eGFR	One Year	Two Year	Five Year	Average
Brier					
Five	< 60	0.74 (0.74, 0.75)	0.72 (0.72, 0.72)	0.67 (0.66, 0.67)	0.69 (0.69, 0.69)
Five	< 30	0.76 (0.75, 0.76)	0.71 (0.71, 0.72)	0.65 (0.65, 0.66)	0.68 (0.68, 0.69)
Three	< 60	0.72 (0.72, 0.73)	0.72 (0.72, 0.72)	0.66 (0.66, 0.67)	0.68 (0.68, 0.69)
Three	< 30	0.72 (0.71, 0.72)	0.74 (0.73, 0.74)	0.67 (0.67, 0.68)	0.69 (0.68, 0.69)
Two	< 60	0.74 (0.74, 0.74)	0.72 (0.72, 0.73)	0.67 (0.67, 0.68)	0.69 (0.68, 0.69)
Two	< 30	0.74 (0.73, 0.74)	0.72 (0.72, 0.73)	0.72 (0.71, 0.72)	0.69 (0.69, 0.69)
c-statistic					
Five	< 60	0.88 (0.88, 0.88)	0.86 (0.85, 0.86)	0.83 (0.83, 0.84)	0.84 (0.84, 0.84)
Five	< 30	0.88 (0.87, 0.88)	0.87 (0.87, 0.87)	0.86 (0.86, 0.86)	0.84 (0.84, 0.85)
Three	< 60	0.87 (0.87, 0.87)	0.87 (0.86, 0.87)	0.84 (0.84, 0.84)	0.84 (0.84, 0.84)
Three	< 30	0.86 (0.86, 0.86)	0.87 (0.87, 0.87)	0.85 (0.85, 0.85)	0.84 (0.84, 0.85)
Two	< 60	0.86 (0.86, 0.86)	0.86 (0.86, 0.87)	0.84 (0.84, 0.84)	0.84 (0.84, 0.84)
Two	< 30	0.86 (0.86, 0.86)	0.87 (0.87, 0.88)	0.81 (0.81, 0.81)	0.84 (0.84, 0.84)
Intercept					
Five	< 60	0.00 (-0.00, 0.00) -0.01 (-0.01, -0.01) -0.01 (-0.01, -0.01) 0.01 (0.00, 0.01) -0.00 (-0.00, 0.00) -0.01 (-0.01, -0.00) -0.02 (-0.02, -0.02)	0.01 (0.00, 0.01) 0.01 (0.01, 0.01) -0.00 (-0.00, 0.00) -0.00 (-0.00, -0.00) 0.00 (0.00, 0.01) 0.00 (0.00, 0.00) 0.00 (0.00, 0.00)	-0.01 (-0.01, -0.00) -0.00 (-0.01, -0.00) -0.01 (-0.01, -0.00) -0.04 (-0.05, -0.04) 0.01 (0.00, 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) -0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) 0.00 (0.00, 0.01) -0.00 (-0.01, -0.00)
Five	< 30	0.00 (0.00, 0.01) -0.00 (-0.01, -0.00) -0.00 (-0.01, -0.00) 0.01 (0.01, 0.01) 0.00 (0.00, 0.00)	0.00 (0.00, 0.00) -0.01 (-0.02, -0.01) -0.02 (-0.02, -0.01) -0.00 (-0.00, -0.00) -0.00 (-0.00, -0.00)	0.00 (0.00, 0.01) -0.00 (-0.00, 0.00) -0.01 (-0.01, -0.01) -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) -0.00 (-0.00, 0.00)
Three	< 60	-0.00 (-0.01, -0.00) 0.00 (-0.00, 0.00) -0.01 (-0.01, -0.00)	0.01 (0.01, 0.02) 0.00 (-0.00, 0.00) 0.00 (0.00, 0.00)	0.02 (0.02, 0.02) -0.02 (-0.02, -0.02) 0.02 (0.02, 0.02)	0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) -0.00 (-0.00, -0.00)
Three	< 30	-0.00 (-0.01, -0.00) 0.00 (-0.00, 0.00) 0.01 (0.00, 0.01)	0.00 (0.00, 0.00) 0.01 (0.00, 0.01) 0.01 (0.01, 0.01)	0.00 (0.00, 0.00) -0.01 (-0.02, -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)
Two	< 60	-0.01 (-0.01, -0.00)	0.01 (0.00, 0.01)	0.00 (-0.00, 0.00)	-0.00 (-0.01, -0.00)
Two	< 30	-0.03 (-0.03, -0.03)	0.02 (0.01, 0.02)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
Slope					
Three	< 60	*1.19*, 0.01, 0.01 -0.01, *1.07*, -0.02 -0.07, 0.00, *1.11*	*1.16*, -0.01, -0.01 -0.00, *1.11*, -0.01 0.01, -0.00, *1.35*	*1.24*, -0.00, 0.00 -0.07, *1.28*, 0.00 -0.03, 0.01, *1.18*	*1.30*, 0.01, -0.00 0.00, *1.29*, -0.00 -0.00, 0.00, *1.32*
Three	< 30	*1.21*, -0.01, 0.01 -0.02, *1.18*, -0.01 0.02, -0.01, *1.15*	*1.12*, 0.01, 0.00 -0.02, *1.27*, 0.00 0.03, 0.04, *1.18*	*1.34*, 0.12, -0.02 0.02, *1.43*, -0.05 0.00, 0.09, *1.19*	*1.27*, 0.00, -0.00 -0.00, *1.28*, 0.00 -0.00, -0.00, *1.29*
Two	< 60	*1.21*	*1.03*	*1.25*	*1.25*
Two	< 30	*1.05*	*1.16*	*1.38*	*1.28*

Table C.9: External Validation of the Five-State Model, results presented as Estimate (95% CI, where possible)

Predicting	eGFR	One Year	Two Year	Five Year	Average
Brier					
Five	< 60	0.70 (0.70, 0.71)	0.72 (0.71, 0.72)	0.64 (0.64, 0.65)	0.63 (0.63, 0.64)
Five	< 30	0.73 (0.72, 0.73)	0.71 (0.70, 0.71)	0.67 (0.67, 0.68)	0.64 (0.64, 0.65)
Three	< 60	0.69 (0.68, 0.69)	0.71 (0.71, 0.71)	0.67 (0.66, 0.67)	0.64 (0.63, 0.64)
Three	< 30	0.69 (0.68, 0.69)	0.68 (0.68, 0.69)	0.63 (0.63, 0.64)	0.63 (0.62, 0.63)
Two	< 60	0.69 (0.68, 0.69)	0.68 (0.68, 0.69)	0.60 (0.60, 0.60)	0.64 (0.63, 0.64)
Two	< 30	0.73 (0.72, 0.73)	0.67 (0.67, 0.67)	0.64 (0.63, 0.64)	0.63 (0.63, 0.64)
c-statistic					
Five	< 60	0.85 (0.85, 0.85)	0.85 (0.84, 0.85)	0.82 (0.82, 0.82)	0.82 (0.81, 0.82)
Five	< 30	0.85 (0.85, 0.85)	0.83 (0.82, 0.83)	0.82 (0.82, 0.82)	0.81 (0.81, 0.82)
Three	< 60	0.83 (0.83, 0.84)	0.83 (0.82, 0.83)	0.81 (0.81, 0.82)	0.81 (0.81, 0.82)
Three	< 30	0.87 (0.87, 0.87)	0.84 (0.84, 0.84)	0.82 (0.82, 0.82)	0.82 (0.81, 0.82)
Two	< 60	0.84 (0.84, 0.85)	0.85 (0.84, 0.85)	0.81 (0.81, 0.81)	0.82 (0.82, 0.82)
Two	< 30	0.84 (0.83, 0.84)	0.84 (0.84, 0.84)	0.81 (0.81, 0.82)	0.82 (0.81, 0.82)
Intercept					
Five	< 60	-0.01 (-0.01, -0.01) -0.02 (-0.02, -0.01) -0.02 (-0.02, -0.01) 0.04 (0.03, 0.04) 0.00 (0.00, 0.00)	0.01 (0.00, 0.01) -0.01 (-0.02, -0.01) -0.01 (-0.02, -0.01) -0.02 (-0.03, -0.02) -0.01 (-0.01, -0.00)	0.00 (0.00, 0.01) -0.00 (-0.01, -0.00) -0.02 (-0.03, -0.02) 0.01 (0.01, 0.02) 0.00 (-0.00, 0.00)	-0.00 (-0.00, 0.00) -0.00 (-0.01, -0.00) 0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) -0.00 (-0.01, -0.00)
Five	< 30	-0.01 (-0.01, -0.00) 0.02 (0.01, 0.02) -0.01 (-0.01, -0.00) 0.02 (0.01, 0.02) -0.02 (-0.03, -0.02) 0.03 (0.02, 0.03) 0.01 (0.01, 0.01)	0.01 (0.00, 0.01) 0.02 (0.02, 0.03) -0.02 (-0.02, -0.01) -0.02 (-0.02, -0.01) -0.00 (-0.00, -0.00) 0.02 (0.02, 0.02) -0.01 (-0.01, -0.00)	0.00 (-0.00, 0.00) 0.03 (0.02, 0.03) 0.00 (0.00, 0.01) -0.02 (-0.02, -0.01) -0.00 (-0.01, -0.00) -0.00 (-0.00, 0.00) 0.02 (0.01, 0.02)	0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) -0.00 (-0.00, -0.00) 0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00) -0.00 (-0.00, -0.00) -0.00 (-0.00, 0.00)
Three	< 60	0.00 (0.00, 0.00) -0.01 (-0.02, -0.01) -0.02 (-0.02, -0.02)	-0.01 (-0.01, -0.00) 0.02 (0.02, 0.03) 0.01 (0.01, 0.02)	0.01 (0.00, 0.01) -0.03 (-0.04, -0.03) -0.02 (-0.03, -0.02)	0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) 0.00 (-0.00, 0.00)
Three	< 30	-0.00 (-0.01, -0.00) -0.02 (-0.02, -0.02) -0.03 (-0.04, -0.03)	0.01 (0.01, 0.02) -0.00 (-0.01, -0.00) -0.00 (-0.01, -0.00)	0.02 (0.02, 0.02) -0.05 (-0.05, -0.04) 0.01 (0.01, 0.01)	-0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00) -0.00 (-0.00, 0.00)
Two	< 60	0.00 (0.00, 0.01)	-0.02 (-0.02, -0.02)	-0.00 (-0.00, 0.00)	0.00 (-0.00, 0.00)
Two	< 30	-0.02 (-0.02, -0.01)	0.02 (0.02, 0.03)	0.01 (0.01, 0.01)	-0.00 (-0.00, 0.00)
Slope					
Three	< 60	*1.01*, -0.00, -0.00 0.01, *1.21*, -0.05 0.08, -0.05, *1.22*	*1.40*, 0.02, -0.02 -0.03, *1.28*, -0.03 -0.07, 0.06, *1.12*	*1.61*, -0.03, 0.04 0.03, *1.76*, 0.00 0.04, -0.00, *1.45*	*1.46*, -0.00, -0.00 0.00, *1.50*, -0.01 -0.00, -0.00, *1.45*
Three	< 30	*1.24*, -0.00, -0.00 -0.02, *1.20*, -0.05 -0.01, -0.07, *1.34*	*1.39*, 0.03, 0.00 -0.04, *1.24*, 0.04 -0.03, -0.01, *1.18*	*1.28*, 0.02, 0.03 -0.09, *1.38*, 0.02 -0.07, -0.00, *1.44*	*1.51*, -0.00, 0.00 0.00, *1.41*, -0.00 -0.00, -0.00, *1.47*
Two	< 60	*1.38*	*1.12*	*1.25*	*1.46*
Two	< 30	*1.26*	*1.28*	*1.35*	*1.55*

Table C.10: Calibration Slope results for both the External and Internal Validation for the Five-State Model

	Internal < 60	Internal < 30
One Year	*1.14*, -0.01, 0.00, 0.02, 0.00, 0.01 -0.00, *1.09*, -0.05, -0.01, 0.01, 0.00 -0.04, 0.06, *1.22*, 0.02, 0.03, 0.00 -0.06, 0.02, 0.01, *1.25*, 0.00, -0.02 -0.02, 0.03, 0.02, -0.02, *1.13*, 0.00 0.05, -0.04, -0.06, -0.02, -0.00, *1.30*	*1.07*, 0.03, 0.02, 0.02, 0.01, -0.00 0.00, *1.18*, 0.01, 0.00, -0.06, 0.01 0.01, 0.00, *1.03*, 0.00, -0.05, 0.00 -0.04, -0.03, -0.03, *1.11*, 0.01, -0.03 0.00, -0.03, -0.02, -0.02, *1.15*, 0.01 0.02, -0.01, 0.00, -0.00, -0.01, *1.11*
Two Year	*1.42*, -0.05, 0.03, -0.01, -0.05, 0.00 0.02, *1.12*, 0.03, -0.01, 0.03, -0.00 0.01, -0.01, *1.22*, -0.00, 0.00, 0.04 -0.03, 0.00, -0.05, *1.21*, -0.01, 0.00 -0.02, -0.03, -0.00, -0.01, *1.39*, -0.00 -0.03, 0.04, 0.04, -0.00, 0.03, *1.15*	*1.14*, -0.02, 0.04, 0.02, 0.00, -0.09 -0.01, *1.10*, 0.00, -0.07, -0.01, -0.00 -0.03, 0.02, *1.12*, 0.03, 0.00, -0.04 0.00, 0.02, -0.03, *1.11*, -0.02, 0.01 -0.00, -0.06, -0.05, 0.02, *1.11*, -0.03 0.07, -0.01, -0.03, 0.02, 0.04, *1.18*
Five Year	*1.22*, -0.00, 0.05, -0.00, -0.05, 0.05 -0.04, *1.11*, -0.01, 0.03, 0.04, 0.03 0.02, -0.03, *1.24*, -0.03, -0.03, 0.01 0.02, -0.03, -0.00, *1.20*, -0.05, 0.01 0.00, 0.08, -0.00, 0.01, *1.25*, -0.06 0.01, -0.00, -0.00, 0.03, -0.05, *1.14*	*1.27*, 0.02, 0.02, -0.03, 0.04, -0.06 -0.00, *1.20*, -0.00, 0.00, 0.04, -0.00 -0.03, -0.02, *1.22*, 0.02, 0.05, 0.03 0.00, -0.02, -0.01, *1.30*, -0.00, 0.04 0.00, -0.05, 0.00, -0.05, *1.31*, 0.01 -0.01, -0.04, -0.05, 0.02, -0.00, *1.19*
Average	*1.28*, 0.00, -0.00, 0.01, 0.00, 0.00 0.00, *1.28*, 0.00, 0.00, -0.00, 0.01 0.00, 0.01, *1.25*, 0.00, -0.00, 0.00 -0.00, 0.00, 0.00, *1.28*, -0.00, 0.00 -0.00, 0.00, -0.00, 0.01, *1.26*, -0.00 -0.00, 0.01, 0.00, -0.00, -0.00, *1.31*	*1.31*, 0.01, -0.00, 0.00, 0.00, 0.00 -0.00, *1.29*, -0.00, 0.00, -0.00, 0.00 -0.00, 0.00, *1.30*, -0.00, -0.00, -0.00 0.00, -0.00, -0.00, *1.28*, 0.00, -0.01 0.00, 0.00, 0.00, 0.00, *1.27*, -0.01 0.01, -0.00, -0.00, -0.00, 0.00, *1.28*

References

[1] Hippocrates and F. Adams, *The genuine works of Hippocrates;* New York, W. Wood and company, 1886.

[2] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study," *BMJ*, vol. 357, May 2017, doi: 10.1136/bmj.j2099.

[3] H. Hemingway *et al.*, "Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes," *BMJ*, vol. 346, p. e5595, Feb. 2013, doi: 10.1136/bmj.e5595.

[4] K. Thygesen, J. S. Alpert, H. 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*, vol. 50, no. 22, pp. 2173–2195, Nov. 2007, doi: 10.1016/j.jacc.2007.09.011.

[5] Probst *et al.*, "Long-Term Prognosis of Patients Diagnosed With Brugada Syndrome," *Circulation*, vol. 121, no. 5, pp. 635–643, Feb. 2010, doi: 10.1161/CIRCULATIONAHA.109.887026.

[6] R. D. Riley, D. van der Windt, P. Croft, and K. G. M. Moons, *Prognosis Research in Healthcare: Concepts, Methods, and Impact*, First. Oxford University Press, 2019.

[7] R. D. Riley *et al.*, "Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research," *PLoS Medicine*, vol. 10, no. 2, Feb. 2013, doi: 10.1371/journal.pmed.1001380.

[8] E. W. Steyerberg *et al.*, "Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research," *PLOS Medicine*, vol. 10, no. 2, p. e1001381, Feb. 2013, doi: 10.1371/journal.pmed.1001381.

[9] P. Royston, K. G. M. Moons, D. G. Altman, and Y. Vergouwe, "Prognosis and prognostic research: Developing a prognostic model," *BMJ*, vol. 338, p. b604, Mar. 2009, doi: 10.1136/bmj.b604.

[10] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. M. Moons, "Prognosis and prognostic research: Validating a prognostic model," *BMJ*, vol. 338, p. b605, May 2009, doi: 10.1136/bmj.b605.

[11] K. G. M. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman, "Prognosis and prognostic research: What, why, and how?" *BMJ*, vol. 338, p. b375, Feb. 2009, doi: 10.1136/bmj.b375.

[12] A. D. Hingorani *et al.*, "Prognosis research strategy (PROGRESS) 4: Stratified medicine research," *BMJ*, vol. 346, p. e5793, Feb. 2013, doi: 10.1136/bmj.e5793.

[13] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement," *BMC Medicine*, vol. 13, no. 1, p. 1, Jan. 2015, doi: 10.1186/s12916-014-0241-z.

[14] K. G. M. Moons *et al.*, "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration," *Annals of Internal Medicine*, vol. 162, no. 1, p. W1, Jan. 2015, doi: 10.7326/M14-0698.

[15] R. D. Riley *et al.*, "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges," *BMJ*, vol. 353, Jun. 2016, doi: 10.1136/bmj.i3140.

[16] G. S. Collins, O. Omar, M. Shanyinde, and L.-M. Yu, "A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods," *Journal of Clinical Epidemiology*, vol. 66, no. 3, pp. 268–277, Mar. 2013, doi: 10.1016/j.jclinepi.2012.06.020.

[17] P. Royston, D. G. Altman, and W. Sauerbrei, "Dichotomizing continuous predictors in multiple regression: A bad idea," *Statistics in Medicine*, vol. 25, no. 1, pp. 127–141, Jan. 2006, doi: 10.1002/sim.2331.

[18] S. E. Bleeker *et al.*, "External validation is necessary in prediction research: A clinical example," *Journal of Clinical Epidemiology*, vol. 56, no. 9, pp. 826–832, Sep. 2003, doi: 10.1016/s0895-4356(03)00207-5.

[19] S. B. Hanauer, "Exploring the controversial themes of IBD," *Inflammatory Bowel Diseases*, vol. 15, no. S1, pp. S1–S10, 2009, doi: 10.1002/ibd.20945.

[20] J. Hippisley-Cox *et al.*, "Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2," *BMJ*, vol. 336, no. 7659, pp. 1475–1482, Jun. 2008, doi: 10.1136/bmj.39609.449676.25.

[21] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation

study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373–1379, Dec. 1996, doi: 10.1016/S0895-4356(96)00236-3.

[22] R. D. Riley *et al.*, "Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes," *Statistics in Medicine*, vol. 38, no. 7, pp. 1276–1296, 2019, doi: 10.1002/sim.7992.

[23] W. Sauerbrei, P. Royston, and H. Binder, "Selection of important variables and determination of functional form for continuous predictors in multivariable model building," *Statistics in Medicine*, vol. 26, no. 30, pp. 5512–5528, 2007, doi: 10.1002/sim.3148.

[24] Y. Vergouwe, E. W. Steyerberg, M. J. C. Eijkemans, and J. D. F. Habbema, "Substantial effective sample sizes were required for external validation studies of predictive logistic regression models," *Journal of Clinical Epidemiology*, vol. 58, no. 5, pp. 475–483, May 2005, doi: 10.1016/j.jclinepi.2004.06.017.

[25] C. Counsell and M. Dennis, "Systematic review of prognostic models in patients with acute stroke," *Cerebrovascular Diseases (Basel, Switzerland)*, vol. 12, no. 3, pp. 159–170, 2001, doi: 10.1159/000047699.

[26] 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*, vol. 70, no. 1, pp. 162–168, Jul. 2000, doi: 10.1016/s0003-4975(00)01387-4.

[27] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study," *BMJ (Clinical research ed.)*, vol. 335, no. 7611, p. 136, Jul. 2007, doi: 10.1136/bmj.39261.471806.55.

[28] J. L. Haybittle *et al.*, "A prognostic index in primary breast cancer." *British Journal of Cancer*, vol. 45, no. 3, pp. 361–366, Mar. 1982.

[29] J. H. Todd *et al.*, "Confirmation of a prognostic index in primary breast cancer," *British Journal of Cancer*, vol. 56, no. 4, pp. 489–492, Oct. 1987, doi: 10.1038/bjc.1987.230.

[30] A. Pate, R. Emsley, D. M. Ashcroft, B. Brown, and T. 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*, vol. 17, no. 1, p. 134, Jul. 2019, doi: 10.1186/s12916-019-1368-8.

[31] P. Bhatnagar, K. Wickramasinghe, J. Williams, M. Rayner, and N. Townsend, "The epidemiology of cardiovascular disease in the UK 2014," *Heart*, vol. 101, no. 15, pp. 1182–1189,

Aug. 2015, doi: 10.1136/heartjnl-2015-307516.

[32] D. A. Jenkins, M. Sperrin, G. P. Martin, and N. Peek, "Dynamic models to predict health outcomes: Current status and methodological challenges," *Diagnostic and Prognostic Research*, vol. 2, no. 1, p. 23, Dec. 2018, doi: 10.1186/s41512-018-0045-2.

[33] B. Liquet, J.-F. Timsit, and V. Rondeau, "Investigating hospital heterogeneity with a multistate frailty model: Application to nosocomial pneumonia disease in intensive care units," *BMC medical research methodology*, vol. 12, p. 79, Jun. 2012, doi: 10.1186/1471-2288-12-79.

[34] K. I. E. Snell *et al.*, "Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model," *Journal of Clinical Epidemiology*, vol. 69, pp. 40–50, Jan. 2016, doi: 10.1016/j.jclinepi.2015.05.009.

[35] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media, 2008.

[36] B. V. Calster, D. Nieboer, Y. Vergouwe, B. D. Cock, M. J. Pencina, and E. W. Steyerberg, "A calibration hierarchy for risk models was defined: From utopia to empirical data," *Journal of Clinical Epidemiology*, vol. 74, pp. 167–176, Jun. 2016, doi: 10.1016/j.jclinepi.2015.12.005.

[37] P. Royston and D. G. Altman, "External validation of a Cox prognostic model: Principles and methods," *BMC Medical Research Methodology*, vol. 13, no. 1, p. 33, Mar. 2013, doi: 10.1186/1471-2288-13-33.

[38] C. S. Crowson, E. J. Atkinson, and T. M. Therneau, "Assessing Calibration of Prognostic Risk Scores," *Statistical methods in medical research*, vol. 25, no. 4, pp. 1692–1706, Aug. 2016, doi: 10.1177/0962280213497434.

[39] H. C. van Houwelingen, "Validation, calibration, revision and combination of prognostic survival models," *Statistics in Medicine*, vol. 19, no. 24, pp. 3401–3415, 2000, doi: 10.1002/1097-0258(20001230)19:24<3401::AID-SIM554>3.0.CO;2-2.

[40] P. Royston, "Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities:" *The Stata Journal*, Dec. 2014, doi: 10.1177/1536867X140140

[41] M. P. Perme and P. K. Andersen, "Checking hazard regression models using pseudo-observations," *Statistics in medicine*, vol. 27, no. 25, pp. 5309–5328, Nov. 2008, doi: 10.1002/sim.3401.

[42] P. 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*, vol. 15, no. 1, pp. 275–291, Apr. 2015, doi: 10.1177/1536867X1501500116.

[43] T. A. Gerds and M. Schumacher, "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times," *Biometrical Journal*, vol. 48,

no. 6, pp. 1029–1040, 2006, doi: 10.1002/bimj.200610301.

[44] C. Spitoni, V. Lammens, and H. Putter, "Prediction errors for state occupation and transition probabilities in multi-state models," *Biometrical Journal. Biometrische Zeitschrift*, vol. 60, no. 1, pp. 34–48, Jan. 2018, doi: 10.1002/bimj.201600191.

[45] X. Han, Y. Zhang, and Y. Shao, "On comparing two correlated C indices with censored survival data," *Statistics in medicine*, vol. 36, no. 25, pp. 4041–4049, Nov. 2017, doi: 10.1002/sim.7414.

[46] X. Liu, Z. Jin, and J. H. Graziano, "Comparing paired biomarkers in predicting quantitative health outcome subject to random censoring," *Statistical methods in medical research*, vol. 25, no. 1, pp. 447–457, Feb. 2016, doi: 10.1177/0962280212460434.

[47] A. Burton, D. G. Altman, P. Royston, and R. L. Holder, "The design of simulation studies in medical statistics," *Statistics in Medicine*, vol. 25, no. 24, pp. 4279–4292, Dec. 2006, doi: 10.1002/sim.2673.

[48] P. K. Andersen and M. Pohar Perme, "Pseudo-observations in survival analysis," *Statistical Methods in Medical Research*, vol. 19, no. 1, pp. 71–99, Feb. 2010, doi: 10.1177/0962280209105020.

[49] Wildscop, "Biostatistics and epidemiology with R: Weighted Logistic Regression in R, SPSS, Stata," *biostatistics and epidemiology with R*. Feb-2013.

[50] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, 2019, doi: 10.1002/sim.8086.

[51] R. C. Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, Vienna,

[52] H. Wickham, "The tidy tools manifesto." https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.htm Nov-2017.

[53] T. Therneau, "A package for survival analysis in R," p. 89, Mar. 2020.

[54] M. P. Perme, M. Gerster, and K. Rodrigues, "Pseudo: Computes Pseudo-Observations for Modeling." Jul-2017.

[55] W. Chang *et al.*, "Shiny: Web Application Framework for R." Mar-2020.

[56] E. S. Johnson, M. L. Thorp, X. Yang, O. L. Charansonney, and D. H. Smith, "Predicting renal replacement therapy and mortality in CKD," *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, vol. 50, no. 4, pp. 559–565, Oct. 2007, doi: 10.1053/j.ajkd.2007.07.006.

[57] M. J. Landray *et al.*, "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*, vol. 56, no. 6, pp. 1082–1094, Dec. 2010, doi: 10.1053/j.ajkd.2010.07.016.

[58] N. Bansal *et al.*, "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*, vol. 10, no. 3, pp. 363–371, Mar. 2015, doi: 10.2215/CJN.04650514.

[59] A. Marks *et al.*, "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*, vol. 30, no. 9, pp. 1507–1517, Sep. 2015, doi: 10.1093/ndt/gfv089.

[60] J. P. Wick *et al.*, "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*, vol. 69, no. 5, pp. 568–575, May 2017, doi: 10.1053/j.ajkd.2016.08.035.

[61] E. S. Johnson, M. L. Thorp, R. W. Platt, and D. 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*, vol. 52, no. 4, pp. 653–660, Oct. 2008, doi: 10.1053/j.ajkd.2008.04.026.

[62] E. B. Schroeder *et al.*, "Predicting 5-Year Risk of RRT in Stage 3 or 4 CKD: Development and External Validation," *Clinical journal of the American Society of Nephrology: CJASN*, vol. 12, no. 1, pp. 87–94, Jun. 2017, doi: 10.2215/CJN.01290216.

[63] S. Kulkarni *et al.*, "Transition probabilities between changing sensitization levels, waitlist activity status and competing-risk kidney transplant outcomes using multi-state modeling," *PLOS ONE*, vol. 12, no. 12, p. e0190277, Dec. 2017, doi: 10.1371/journal.pone.0190277.

[64] J. Floege *et al.*, "Development and validation of a predictive mortality risk score from a European hemodialysis cohort," *Kidney International*, vol. 87, no. 5, pp. 996–1008, May 2015, doi: 10.1038/ki.2014.419.

[65] A. C. Hemke, M. B. Heemskerk, M. van Diepen, W. Weimar, F. W. Dekker, and A. J. Hoitsma, "Survival prognosis after the start of a renal replacement therapy in the Netherlands: A retrospective cohort study," *BMC Nephrology*, vol. 14, p. 258, Nov. 2013, doi: 10.1186/1471-2369-14-258.

[66] X.-Y. Cao *et al.*, "Predicting one-year mortality in peritoneal dialysis patients: An analysis of the China Peritoneal Dialysis Registry," *International Journal of Medical Sciences*, vol.

12, no. 4, pp. 354–361, 2015, doi: 10.7150/ijms.11694.

[67] N. Tangri *et al.*, "A predictive model for progression of chronic kidney disease to kidney failure," *JAMA*, vol. 305, no. 15, pp. 1553–1559, Apr. 2011, doi: 10.1001/jama.2011.451.

[68] J. Roy *et al.*, "Statistical Methods for Cohort Studies of CKD: Prediction Modeling," *Clinical journal of the American Society of Nephrology: CJASN*, vol. 12, no. 6, pp. 1010–1017, Jun. 2017, doi: 10.2215/CJN.06210616.

[69] N. Tangri *et al.*, "A Dynamic Predictive Model for Progression of CKD," *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, vol. 69, no. 4, pp. 514–520, Apr. 2017, doi: 10.1053/j.ajkd.2016.07.030.

[70] M. G. Shlipak *et al.*, "Cardiovascular mortality risk in chronic kidney disease: Comparison of traditional and novel risk factors," *JAMA*, vol. 293, no. 14, pp. 1737–1745, Apr. 2005, doi: 10.1001/jama.293.14.1737.

[71] D. E. Weiner *et al.*, "The Framingham predictive instrument in chronic kidney disease," *Journal of the American College of Cardiology*, vol. 50, no. 3, pp. 217–224, Jul. 2007, doi: 10.1016/j.jacc.2007.03.037.

[72] J. J. V. McMurray *et al.*, "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*, vol. 162, no. 4, pp. 748–755.e3, Oct. 2011, doi: 10.1016/j.ahj.2011.07.016.

[73] M. E. Grams and J. Coresh, "Assessing risk in chronic kidney disease: A methodological review," *Nature Reviews. Nephrology*, vol. 9, no. 1, pp. 18–25, Jan. 2013, doi: 10.1038/nrneph.2012.248.

[74] N. Tangri *et al.*, "Risk prediction models for patients with chronic kidney disease: A systematic review," *Annals of Internal Medicine*, vol. 158, no. 8, pp. 596–603, Apr. 2013, doi: 10.7326/0003-4819-158-8-201304160-00004.

[75] C. L. Ramspek, P. W. Voskamp, F. J. van Ittersum, R. T. Krediet, F. W. Dekker, and M. van Diepen, "Prediction models for the mortality risk in chronic dialysis patients: A systematic review and independent external validation study," *Clinical Epidemiology*, vol. 9, pp. 451–464, 2017, doi: 10.2147/CLEP.S139748.

[76] W. Bouwmeester *et al.*, "Reporting and Methods in Clinical Prediction Research: A Systematic Review," *PLOS Medicine*, vol. 9, no. 5, p. e1001221, May 2012, doi: 10.1371/journal.pmed.1001221.

[77] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. 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*, vol. 22, no. 4, pp. 872–880, Jul. 2015, doi: 10.1093/jamia/ocv024.

[78] 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*, vol. 33, no. 2, pp. 298–306, Feb. 2013, doi: 10.1177/0272989X12466731.

[79] A. M. Allen, W. R. Kim, T. M. Therneau, J. J. Larson, J. K. Heimbach, and A. D. Rule, "Chronic kidney disease and associated mortality after liver transplantation–a time-dependent analysis using measured glomerular filtration rate," *Journal of Hepatology*, vol. 61, no. 2, pp. 286–292, Aug. 2014, doi: 10.1016/j.jhep.2014.03.034.

[80] M. E. Grams *et al.*, "Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate," *Kidney International*, vol. 93, no. 6, pp. 1442–1451, Jun. 2018, doi: 10.1016/j.kint.2018.01.009.

[81] D. G. Altman, "Problems in dichotomizing continuous variables," *American Journal of Epidemiology*, vol. 139, no. 4, pp. 442–445, Feb. 1994, doi: 10.1093/oxfordjournals.aje.a117020.

[82] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher, "Dangers of Using 'Optimal' Cutpoints in the Evaluation of Prognostic Factors," *JNCI: Journal of the National Cancer Institute*, vol. 86, no. 11, pp. 829–835, Jun. 1994, doi: 10.1093/jnci/86.11.829.

[83] D. G. Altman and P. Royston, "The cost of dichotomising continuous variables," *BMJ*, vol. 332, no. 7549, p. 1080, May 2006, doi: 10.1136/bmj.332.7549.1080.

[84] C. Bennette and A. Vickers, "Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents," *BMC Medical Research Methodology*, vol. 12, no. 1, p. 21, Feb. 2012, doi: 10.1186/1471-2288-12-21.

[85] M. M. Butts and T. W. H. Ng, "Chopped liver? OK. Chopped data? Not OK," in *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, New York, NY, US: Routledge/Taylor & Francis Group, 2009, pp. 361–386.

[86] P. M. Cumberland *et al.*, "Ophthalmic statistics note: The perils of dichotomising continuous variables," *British Journal of Ophthalmology*, vol. 98, no. 6, pp. 841–843, Jun. 2014, doi: 10.1136/bjophthalmol-2014-304930.

[87] N. V. Dawson and R. Weiss, "Dichotomizing continuous variables in statistical analysis: A practice to avoid," *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, vol. 32, no. 2, pp. 225–226, doi: 10.1177/0272989X12437605.

[88] T. E. Dinero, "Seven reasons why you should not categorize continuous data," *Journal of Health & Social Policy*, vol. 8, no. 1, pp. 63–72, 1996, doi: 10.1300/J045v08n01_06.

[89] J. R. Irwin and G. H. McClelland, "Negative Consequences of Dichotomizing Continuous Predictor Variables," *Journal of Marketing Research*, vol. 40, no. 3, pp. 366–371, Aug. 2003, doi: 10.1509/jmkr.40.3.366.19237.

[90] O. Kuss, "The danger of dichotomizing continuous variables: A visualization," *Teaching Statistics*, vol. 35, no. 2, pp. 78–79, 2013, doi: 10.1111/test.12006.

[91] K. Metze, "Dichotomization of continuous data–a pitfall in prognostic factor studies," *Pathology, Research and Practice*, vol. 204, no. 3, pp. 213–214, 2008, doi: 10.1016/j.prp.2007.12.002.

[92] O. Naggara, J. Raymond, F. Guilbert, D. Roy, A. Weill, and D. G. Altman, "Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms," *AJNR. American journal of neuroradiology*, vol. 32, no. 3, pp. 437–440, Mar. 2011, doi: 10.3174/ajnr.A2425.

[93] S. V. Owen and R. D. Froman, "Why carve up your continuous data?" *Research in Nursing & Health*, vol. 28, no. 6, pp. 496–503, 2005, doi: 10.1002/nur.20107.

[94] J. M. Schellingerhout, M. W. Heymans, H. C. W. de Vet, B. W. Koes, and A. P. Verhagen, "Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain," *Journal of Clinical Epidemiology*, vol. 62, no. 8, pp. 868–874, Aug. 2009, doi: 10.1016/j.jclinepi.2008.10.010.

[95] D. L. Streiner, "Breaking up is hard to do: The heartbreak of dichotomizing continuous data," *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, vol. 47, no. 3, pp. 262–266, Apr. 2002, doi: 10.1177/070674370204700307.

[96] C. van Walraven and R. G. Hart, "Leave 'em alone - why continuous variables should be analyzed as such," *Neuroepidemiology*, vol. 30, no. 3, pp. 138–139, 2008, doi: 10.1159/000126908.

[97] A. M. Vintzileos, Y. Oyelese, and C. V. Ananth, "The "anathema" of arbitrary categorization of continuous predictors," *American Journal of Obstetrics and Gynecology*, vol. 210, no. 3, pp. 200–203, Mar. 2014, doi: 10.1016/j.ajog.2013.09.042.

[98] C. R. Weinberg, "How Bad Is Categorization?" *Epidemiology*, vol. 6, no. 4, pp. 345–347, 1995.

[99] M. van Smeden, T. L. Lash, and R. H. H. Groenwold, "Reflection on modern methods: Five myths about measurement error in epidemiological research," *International Journal of Epidemiology*, Oct. 19AD, doi: 10.1093/ije/dyz251.

[100] J. Sun, "Interval Censoring," in *Encyclopedia of Biostatistics*, American Cancer Society,

2005.

[101] R. A. Hoefield *et al.*, "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*, vol. 56, no. 6, pp. 1072–1081, Dec. 2010, doi: 10.1053/j.ajkd.2010.06.010.

[102] R. Chinnadurai, C. Chrysochou, and P. A. Kalra, "Increased Risk for Cardiovascular Events in Patients with Diabetic Kidney Disease and Non-Alcoholic Fatty Liver Disease," *Nephron*, vol. 141, no. 1, pp. 24–30, 2019, doi: 10.1159/000493472.

[103] A. S. Levey *et al.*, "A new equation to estimate glomerular filtration rate," *Annals of Internal Medicine*, vol. 150, no. 9, pp. 604–612, May 2009, doi: 10.7326/0003-4819-150-9-200905050-00006.

[104] J. P. New, N. D. Bakerly, D. Leather, and A. Woodcock, "Obtaining real-world evidence: The Salford Lung Study," *Thorax*, vol. 69, pp. 1152–1154, 2014, doi: http://dx.doi.org/10.1136/thoraxjnl-2014-205259.

[105] K. Matsushita *et al.*, "Cohort Profile: The Chronic Kidney Disease Prognosis Consortium," *International Journal of Epidemiology*, vol. 42, no. 6, pp. 1660–1668, Dec. 2013, doi: 10.1093/ije/dys173.

[106] L. G. Forni *et al.*, "Renal recovery after acute kidney injury," *Intensive Care Medicine*, vol. 43, no. 6, pp. 855–866, 2017, doi: 10.1007/s00134-017-4809-x.

[107] "KDIGO Clinical Practice Guideline for Acute Kidney Injury," *OFFICIAL JOURNAL OF THE INTERNATIONAL SOCIETY OF NEPHROLOGY*, p. 141, 2012.

[108] S. van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, no. 1, pp. 1–67, Dec. 2011, doi: 10.18637/jss.v045.i03.

[109] C. Jackson, "Flexsurv: A Platform for Parametric Survival Modelling in R," p. 33.

[110] B. Ripley and W. Venables, "Package 'nnet'," Feb-2016.

[111] D. Vaughan and M. Dancho, "Furrr: Apply Mapping Functions in Parallel using Futures." May-2018.

[112] P. Royston and M. 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*, vol. 21, no. 15, pp. 2175–2197, Aug. 2002, doi: 10.1002/sim.1203.

[113] K. G. M. Moons, D. G. Altman, Y. Vergouwe, and P. Royston, "Prognosis and prognostic research: Application and impact of prognostic models in clinical practice," *BMJ*, vol. 338, p. b606, Jun. 2009, doi: 10.1136/bmj.b606.

[114] G. Venkat-Raman *et al.*, "New primary renal diagnosis codes for the ERA-EDTA," *Nephrology Dialysis Transplantation*, vol. 27, no. 12, pp. 4414–4419, Dec. 2012, doi: 10.1093/ndt/gfs461.

[115] C. P. Kovesdy *et al.*, "Past Decline Versus Current eGFR and Subsequent ESRD Risk," *Journal of the American Society of Nephrology*, vol. 27, no. 8, pp. 2447–2455, Aug. 2016, doi: 10.1681/ASN.2015060687.

[116] D. M. J. Naimark *et al.*, "Past Decline Versus Current eGFR and Subsequent Mortality Risk," *Journal of the American Society of Nephrology*, vol. 27, no. 8, pp. 2456–2466, Aug. 2016, doi: 10.1681/ASN.2015060688.

[117] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen, "Multi-state models for the analysis of time-to-event data," *Statistical methods in medical research*, vol. 18, no. 2, pp. 195–222, Apr. 2009, doi: 10.1177/0962280208092301.

[118] H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in biostatistics: Competing risks and multi-state models," *Statistics in Medicine*, vol. 26, no. 11, pp. 2389–2430, May 2007, doi: 10.1002/sim.2712.

[119] A. M. Wood, I. R. White, and P. Royston, "How should variable selection be performed with multiply imputed data?" *Statistics in Medicine*, vol. 27, no. 17, pp. 3227–3246, Jul. 2008, doi: 10.1002/sim.3177.

[120] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc, 1984.

[121] I. R. White and P. Royston, "Imputing missing covariate values for the Cox model," *Statistics in Medicine*, vol. 28, no. 15, pp. 1982–1998, Jul. 2009, doi: 10.1002/sim.3618.

[122] K. J. M. Janssen *et al.*, "Dealing with missing predictor values when applying clinical prediction models," *Clinical Chemistry*, vol. 55, no. 5, pp. 994–1001, May 2009, doi: 10.1373/clinchem.2008.115345.

[123] M. Schomaker and C. Heumann, "Bootstrap inference when using multiple imputation," *Statistics in Medicine*, vol. 37, no. 14, pp. 2252–2266, 2018, doi: 10.1002/sim.7654.

[124] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[125] B. V. Calster, V. V. Belle, Y. Vergouwe, D. Timmerman, S. V. Huffel, and E. W. Steyerberg,

"Extending the c-statistic to nominal polytomous outcomes: The Polytomous Discrimination Index," *Statistics in Medicine*, vol. 31, no. 23, pp. 2610–2626, 2012, doi: 10.1002/sim.5321.

[126] K. V. Hoorde, Y. Vergouwe, D. Timmerman, S. V. Huffel, E. W. Steyerberg, and B. V. Calster, "Assessing calibration of multinomial risk prediction models," *Statistics in Medicine*, vol. 33, no. 15, pp. 2585–2596, Jul. 2014, doi: 10.1002/sim.6114.