Multi-State Clinical Prediction Models in Renal Replacement Therapy

A Thesis
Presented to
The Division of Division of Informatics, Imaging and Data Science
University of Manchester

In Partial Fulfillment
of the Requirements for the Degree
PhD Medicine

Michael Andrew Barrowman

March 2020

Approved for the Division
(School of Health Sciences)

Dr. Matthew Sperrin, Prof. Niels Peek, Dr. Glen Martin, Dr. Mark Lambie

Acknowledgements

I would like to thank my…

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Introduction

Overall introduction to the Thesis. This doesn't seem to want to show up.

Why is this?

0.1 Subsectioing

Testing

Chapter 1

Literature Report

This is the first chapter of my thesis and will include a brief summary of what the current literature looks like. It will be split into sections and subsections as specified in my ToDo List

1.1 Introduction

lorem ipsum blah blah blah

1.2 Clinical Prediction Models

The idea of prognosis dates back to ancient Greece with the work of Hippocrates [**Cite: Hippocrates**] and is derived from the Greek for "know before" meaning to forecast the future. Within the sphere of healthcare…

Prognosis research can be broken down into four main categories (with three subcategories [1]:

- Type I: Fundamental prognosis research [**Cite: LR:1**]
- Type II: Prognostic factor research[**Cite: LR:2**]
- Type III: Prognostic model research[**Cite: LR:3**]
  - Model development[**Cite: LR:34**]
  - Model validation[**Cite: LR:35**]
  - Model impact evaluation[**Cite: LR:36**]
- Type IV: Stratified Medicine [**Cite: LR:4**]

1.2.1 Fundamental Prognosis Research

What is it?

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.

1.2.3 Prognostic Model Research

Model Development

Model Validation

Impact Evaluation

1.2.4 Stratified Medicine

1.2.5 Examples

1.3 Competing Risks & Multi-State Models

Chapter 2

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

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

$$A = \pi r^2$$

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

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. 1. 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:

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

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

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

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].

[**Insert Figure 1**]

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. 2, where T is the event time and δ is the event type indicator (1 for event-of-interest and 2 for competing event).

Chapter 4

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

4.1 Introduction

Clinical prediction models (CPMs) need to be validated before they are used. 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 [1]. The simplest assessment of calibration is the calibration-in-the-large, which tests for agreement in mean calibration (the weakest form of calibration) [2]. With continuous or binary outcomes, such a test is straightforward: 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. 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 [3]. 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. The lack of intercept can be overcome provided sufficient information concerning the baseline survival curve is available (although this is rarely the case [4]). 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 [3], [5]. 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 [6] presents an alternative approach for calibration at external validation. He uses the approach of pseudo-observations, as described by Perme and Anderson [7] 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 [8] 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. QRISK used the overall KM approach in the 2007 paper [5] with good results (6.34% predicted vs 6.25% observed in women and 8.86% predicted vs 8.88% observed in men), but bad results in the QRISK3 update [9] (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 18 years) and an important change between the update was the lower age limit moved from 35 to 25. 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. 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 [10] 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 [11] who demonstrated that any proper scoring rule can be improved by the use of the IPCW. This work has been added to by Han et al [12] and Liu et al [13] who demonstrated that the c-statistic is also suitable. 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 involved from Matt**]

4.2.2 Aims

The aim of this study is to formalise 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

4.2.4 Methods

4.2.5 Estimands

4.2.6 Performance Measures

4.3 Results

4.4 Discussion

Weighting = Good.

Not Weighting = Bad.

Chapter 5

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

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

6.1 Introduction

Chronic Kidney Disease (CKD) is a progressive disease that affects the ability of the kidneys to filter toxins from the blood. Patients with End-stage Renal Disease (ESRD) are treated using Renal Replacement Therapy (RRT), which collectively describes the treatments designed to emulate the processes performed by the failing kidneys. The three most common treatment modalities are haemodialysis (HD), peritoneal dialysis (PD) and kidney transplant (Tx). The more severe stages of CKD (stages 3 - 5) affects approximately 2.6 million people over the age of 16 in England [1] with around 63 thousand adult patients registered for RRT in 2015 [2], of which 8 thousand were new patients [3]. Previous prognostic models have been developed to predict mortality [4]–[8], ESRD [5], the commencements of RRT [7], [9]–[11] or mortality after beginning dialysis [12]–[14]. Some previous models have used the commencement of RRT as a proxy for ESRD [15]–[17], while others have investigated the occurrence of cardiovascular events, which are common amongst CKD patients [18]–[20]. Reviews by Grams & Coresh [21], Tangri et al [22] and Ramspek et al [23], 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 [24], [25]. Methodologically, the majority of existing CKD prediction models fail to account for completing events [6], [8], [26], have high risks of bias [4], [5], [9] or are otherwise flawed compared to modern clinical prediction standards [24], [27]. In 2013, Begun et al [28] developed a Multi-State Clinical Model for assessing patient progression through the severity stages of CKD (III-V), RRT and/or death. In 2014, Allen et al [29] focused a similar model on liver transplant patients. In 2017, Kulkarni et al [11] developed an MSM focusing on the categories of Calculated Panel Reactive Antibodies and transplant and/or death. Most recently, in 2018, Grams et al [30] 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 [31]–[49]. 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 information when done before the data analysis stage and so these models go against the current statistical recommendations (cite). These kinds of assumptions are also subject to measurement error and interval censoring (cite), 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 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 papers have undergone any validation process, whether internal or external [50]. It is also important to note that although these models can be used to predict patient outcomes, these models 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 [51]. The fourth model (Grams et al), 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 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. We also model using extensions to traditional survival analysis to incorporate heterogeneity within the population as much as possible and to allow for the prediction of patient outcomes at any future time point (within the time-scales of the study). 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.

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 [52], [53] 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/1.73m2 (calculated using the CKD-EPI formula [54]) 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) [55]. eGFR, uPCR, comorbidity and blood results were measured longitudinally throughout patients time within the cohort. Patient start dates for our model was assigned as their first date after their consent date at which their eGFR was recorded to be below 60ml/min/1.73m2. 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/1.73m2 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 [56] 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 [30]. In SKS, start dates were chosen to be the first date after consent where their eGFR was recorded to be less than 60ml/min/1.73m2.

In SERPR, start dates were calculated by removing the first recorded eGFR measurement for all patients, any eGFR measurements from before a patient turned 18 and any during an AKI episode [57]. 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 [58]. 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 [59]. All analysis was done in R 3.6.2 [60] using the various tidyverse packages [61], as well as the mice [62], flexsurv [63], nnet [64] and furrr [65] packages.

6.2.2 Development

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 1). The Two-State model was analogous to 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). Variables considered as covariates were demographics (sex, age, smoking status and alcohol consumption), comorbidities (congestive cardiac failure (CCF), chronic obstructive pulmonary disease (COPD), cerebrovascular accident (CVA), hypertension (HT), diabetes mellitus (DM), ischemic heart disease (IHD), chronic liver disease (LD), myocardial infarction (MI), peripheral vascular disease (PVD) and slid tumour (ST)), physical parameters (BMI, blood pressure), blood results (haemoglobin, albumin, calcium and phosphate measures), urine protein creatinine ratio (uPCR) and primary renal diagnosis (grouped as per ERA-EDTA classifications [66]). Ethnicity was assessed in the populations, but due to extreme homogeneity, it was omitted as a potential predictor from the models. uPCR and eGFR Rate of change were also calculated [67], [68] as the difference between the two most recent measures divided by time difference in days. log(Age) was considered as a covariate and then Age was centred at 60 years and squared to account for the variety of effects that Age can have on the transitions involved. To account for any time trend in overall change in treatment techniques, log(-Calendar Time) were included as a covariates (cite). Calendar Time was defined as time of study entry minus 1st January 2019, ensuring CalendarTime is always negative so that patients who entered the study longer ago have a higher value for this covariate. Each transition was modelled under a proportional hazards assumption using the Royston-Parmar technique [69] 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 [70]. For variable selection, we stacked the imputed datasets together to create a larger, pseudo-population [71] and performed backwards-forwards selection based on minimising the AIC at each step (cite). This was repeated for each transition and for different numbers of evenly spaced knots, K={0,1,2,3,4,5}, which 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 [72]. This gave a model fully defined by smooth cubic splines representing the cumulative cause-specific hazard and individualised proportional hazards for each transition.

6.2.3 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. 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 < 30ml/min/1.73m2. 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 1-year, 2-years, 5-years and 10-years (although 10-year validation is not possible for the external validation as the longest follow-up in the validation dataset was 9.5 years). As well as presenting the performance measures over time. The overall accuracy of each model was assessed using the MSM adjusted Brier Score [self-cite], which is a proper score function 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 [73] [self-cite]. 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) [74] [self-cite], which extends the logistic regression to three or more mutually exclusive outcomes [75]. This produces an intercept vector of length K-1 and a Slope-matrix of dimension (K-1) x (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.

6.3 Results

6.3.1 Data Sources

As seen in table 1, 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 (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). 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.

[**Insert Table One (continuous)**]

Table 2 shows a breakdown of the categorical variables across the populations. In the development population, there are far more males than females, which is actually against what is believed to occur in the general public (cite), whereas in the validation population the proportions are much more matched. Ethnicity was very homogeneous in the SKS dataset, and 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).

[**Insert Table One (categorical)**]

Overall, there were high levels of comorbidities within the SKS population as shown in table 3, 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.

[**Insert Table One (comorbs)**]

6.3.2 Development

Table 4shows the full results from the Three-State Models, the results for the Two-State and Five-State Models can be seen in Supplementary Material. Older patients are more likely to transition to the Dead state, regardless of whether from CKD or an RRT state and Older patients were less likely to transition into the RRT states. Increased rates of decline of eGFR had large effects on the transition from CKD into RRT, due to an increase in rates to HD and the high prevalance of transitions to HD since the effect on transitions into PD were relatively low, and eGFR decline actually reduced the rate of transition to Tx.

[**Insert Model Results**]

Female patients were more likely to remain in the CKD state than Males, or to remain in the RRT state once there. Smokers were more likely than Non-/Former Smokers to undergo any transition, apart from CKD to Tx. Blood results had effects on all transitions in some way, and Primary Renal Diagnosis were strongly related to predictions. Patients with Liver Disease were much more likely to undergo a Transplant than those without Liver Disease ($\beta = 12.4$).

6.3.3 Validation

6.4 Discussion

Chapter 7

mscpm Package

Here is where the Vignette for the `\{mscpm\}` package will go, when I get to coding/writing it all up.

It may be easier to write a bit of code to extract the vignette from the package folder (or write something to push it from there to here)

Chapter 8

Conclusion

Here is where my concluding section will go.

The final word.

The end.

References

[1] 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.