Multi-state Clinical Prediction Models in Renal Replacement Therapy

Literature Report

Michael Barrowman

Supervisory Team

Dr. Matthew Sperrin, Dr. Niels Peek, Dr. Mark Lambie

Monday 9th January, 2016

Contents

1 Introduction

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

- Kidney Disease
- Prognostic Research
- Competing Risks and Multi-State Models

The first section, Kidney Disease, will discuss methods of modality of Renal Replacement Therapy (RRT). RRT consists of three treatment methods designed to replace the functionality of the natural kidneys in patients. Haemodialysis (HD) involves the mechanical filtration of patient's blood by removing it from the patient's body, running it through a haemodialysis machine which filters the blood and returns it to the body. Haemodialysis machines can use a variety of methods to filter the blood, which are beyond the scope of this research project, however there are various methods of administering HD through the use of Central Venous Catheter (CVC), ArterioVenous Fistula (AVF) or ArterioVenous Graft (AVG) which will also be discussed as different methods of HD administration can be relevant to patient mortality. Peritoneal Dialysis (PD) involves filling the abdominal cavity with a dialysate fluid which allows substances dissolved in the blood (such as urea, electrolytes, etc), which would usually be filtered by the kidneys, to pass across the membrane of the peritoneum. The waste fluid then needs to be replaced with fresh dialysate regularly. The final modality of RRT is kidney transplantation which, as the name suggests, involves the transplantation of a new kidney from a donor which is expected to replace the old kidney in the most direct way.

The second section discusses prognostic research from a general standpoint. The PROGRESS Series[1–4] classified prognostic research into four main categories[1], however the main focus of this paper will be the third of these, "Prognostic Model Research"[3] as it is the most relevant to the overarching project. The final section will discuss Competing Risks (CRs) and Multistate Models (MSMs), which are extensions of survival analysis. In ordinary survival analysis, patients can be in only one of two states, alive or dead (or their analogues), these extensions can include multiple death states (CR) and/or multiple living states (MSM) giving a more granular and interesting assessment of a patient's journey.

Throughout this paper, examples will be taken from a variety of clinical sources. The two main examples of prognostic models that will be referenced throughout the paper will be the Nottingham Prognostic Index (NPI), originally developed by Haybittle et al[5], and QRISK, developed by Hippisley-Cox et al[6] as a part of the QRESEARCH Group. NPI is a simple and easy-to-use model intended to assist in the prognosis of patients following breast cancer surgery. It uses only 3 variables and so it can be easily calculated by patients and clinicians. QRISK, which was later updated to QRISK2 followed by its subsequent annual updates[7–9] assess patients on their risk of cardiovascular disease, it uses many more variables and so an online calculator (as well as an app and publicly available algorithms) were developed to improve the usability of the more complicated formula[9,10].

2 Kidney Disease

Chronic diseases, especially non-communicable ones, have now become the major cause of morbidity and mortality around the world[11]. In particular, Chronic Kidney Disease (CKD) is a global health concern[12] and is thus a major burden on healthcare utilisation worldwide[13]. This is unsurprising given that, in the UK, 7,411 patients commenced RRT in 2004 alone which equates to a rate of 115 per million people[14]. Part of this prevalence is believed to be increasing due to increased incidences of diabetes[12] which contribute 26.9% of new RRT diagnosis in the UK in 2014[14]. In 2013, the NHS spent 2% of its budget on kidney replacement therapy[12,15] and in 2008, 5.9% of the Medicare expenditure was spent on managing patients with End-Stage Renal Disease (ESRD)[12,16]. The progression of CKD amongst sufferers is believed to be homogeneous with respect to time[17], meaning that it increases continuously at steady rate.

CKD treatment typically consists of either palliative care or a type of RRT. In the real world, it is difficult to make decisions on RRT for patients suffering from ESRD since, as with any disease, there is a lot of variability in the individuals[18]. This variability is particularly prominent amongst older patients, which leads to variation in treatment methods from different physicians[19]. Because of this, it is important to identify, as early as possible, patients who are likely to progress from CKD to ESRD[12]. Tamura et al[18] provides a framework for deciding which RRT patients should receive based on three factors: life expectancy, risks and benefits of competing treatment strategies and patient preference. This framework does not require precision, but rather a general idea of whether

a patient is above or below average (median). These three factors allows for three key choices to be made for the patient: choice of dialysis modality (i.e. HD vs PD), choice of vascular access for HD, and whether or not to be referred for kidney transplantation.

Transferring from one dialysis modality to the other initially increases burden on patients and, for the first few weeks, has a higher mortality rate[20]. Before beginning HD, patients and physicians must decide on the vascular access method which is basically how the HD will be administered. There are three main methods of vascular access: CVCs, AVFs and AVGs[18]. In the US, 80% of patients who are given HD, begin it with a CVC[21]. CVCs are usually used as a temporary placement until a more permanent fistula or graft can be given to the patient[18]. However, it can take time for AVF and AVG patency to occur and so their effects are not immediate. Current guidelines recommend using AVFs over AVGs as the method of permanent access which are both preferred over the temporary access provided by CVC, unless HD is predicted to be only a short-term treatment (i.e. because of expected kidney transplant or extremely high expected mortality)[22]. It is clear that these mortality estimates of patients are currently wildly incorrect as it has been found that two-thirds of deceased patients who had undergone AVF placement had died before it was even used[23].

It is suggested that PD gives an early benefit over HD using CVC due to the high infection rates caused by CVC. However, this benefit might be balanced out by the higher risk of modality failure and a common need to transfer to HD later, which merely pushes the higher CVC risk back[24]. In recent years, It has been observed that survival amongst patients given PD has increased to levels similar to HD[25], although this is likely biased due to the difference in patient's selected for the each modality[18].

Kidney Transplants are often hard to come by as there can be difficulties in finding compatible donors[26]. Living donors provide a better prognosis for recipients than deceased ones, but even deceased-donor transplantation implies a 48-82% decrease in mortality compared to remaining on dialysis[27,28]. For each patient, donors can be classified as being from a Standard Criteria Donor (SCD) or an Expanded Criteria Donor (ECD) list[29]. An ECD is, as the name implies, a much broader list of patients than appear on an SCD. Using an ECD comes with a shorter time on the waiting list for a transplant, but a higher risk of allograft loss and so a decision must be made about a patient of whether they are at higher risk of mortality if they remain on the waiting list for a longer period of time, or whether the risk of an unsuccessful transplant is worth it[18]. As with transfering between dialysis modalities, there is an extremely high increase in risk for the first two weeks after transplantation (compared with staying on dialysis), this risk reduces until 7-8 months after transplantation, where the cumulative mortality of both options becomes equivalent, and afterwards is lower for the transplanted patients[18]. It is worth noting that there is no upper age limit on kidney transplantation[18] and it has actually been found that kidney transplantation was cost-effective amongst patients over 65[30]. This makes sense as, for patient's over 65, the average time spent on the waiting list for a new kidney is 7-8 months[18].

In the UK in 2014, 71.8% of RRT patients had begun with HD, 20.0% were given PD, 8.2% were lined up to receive a kidney transplant[14]. Of the patients who were initially assigned to received HD in 2009, 54.4% had died by 2014 and 34.4% of those still alive had been transfered to a different modality, PD had a lower mortality rate, 35.1%, but a higher transfer rate, 75.3%[14]. These transitions are demonstrated graphically in Figure 1. Although these numbers do not account for the differences between the patients these two modalities were given to, it shows that there are major differences between modalities and that transitioning between treatments is common. These differences between modalities and the prevalence of transitions line up quite well with the idea of using an MSM as a representation for this process.

3 Prognostic Research

The idea of prognosis dates back to ancient Greece with the work of Hippocrates[31] and comes from the Greek for "know before" which means forecasting the future. It is defined as the risk of future health outcomes in patients, particularly patients with a certain disease or health condition. Prognosis allows clinicians to provide patients with a prediction of how their disease will progress and is usually given as a probability of having an event in a prespecified number of years. For example, QRISK2[9] provides a probability that a patient will have a heart attack or stroke in the next 10 years. Prognostic research encompassess any work which enhances the field of prognosis, whether through methodological advancements, specific prognostic models 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, beause of this, most of the key points in the field of 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 effect of screening for asymptomatic diseases (such as with

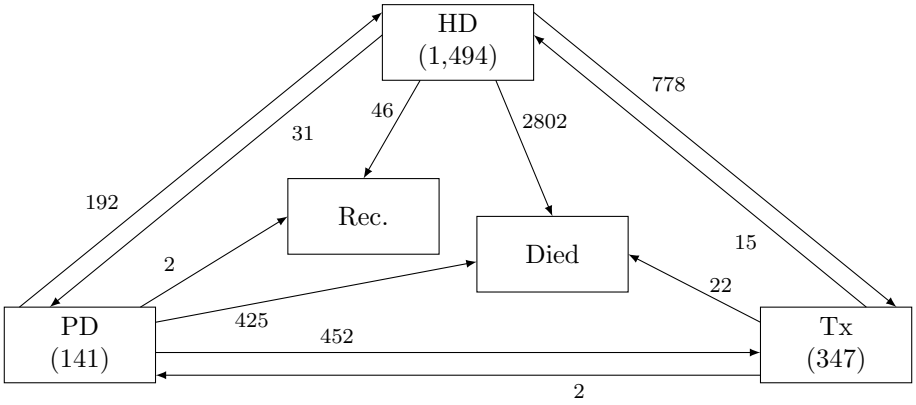

Figure 1: Changes in RRT modality of patients over five years from start of RRT in 2009 to 2014. Tx = Transplant, Rec.= Recovered/Discontinued. Numbers in boxes represent patients still on the same treatment after 5 years, numbers on arrows represent the number of patients transitioning out of that state. Data taken from UK Renal Registry 18th Annual Report[14]

mammograms)[1]. Prognosis research can be used to develop new definitions of diseases, whether a redefinition of an existing disease (such as the extension to the definition of a myocardial infarction to include non-fatal events)[32], or a previously unknown subtype of a disease (such as Brugada syndrome as a type of cardiovascular disease)[33].

Prognostic research can be broken down into four main categories with three subcategories:

- Fundamental prognosis research[1]
- Prognostic factor research[2]
- Prognostic model research:[3]
  - Model development[34]
  - Model validation[35]
  - Model impact evaluation[36]
- Stratified medicine[4]

The model development and validation can sometimes be combined into a single paper[37,38]. Studies into all three of the subcategories of prognostic model research *should* be completed before a model is used in clinical practice[39], although this does not always occur[3]. External validation is considered by some to be more important than the actual derivation of the model as it demonstrates generalisability of the model[12].

Fundamental prognostic research is the study of the underlying methods and reasoning behind prognostic research and includes the methods involved in the other six categories. Projects which are classified as "Fundamental Prognostic Research" can be designed to enhance the abstract theory of prognostic research with little real world application. The "Prognosis and Prognostic Research" series[34–36,40] in the BMJ is a prime example of fundamental prognostic research in so much as it is a general overview of *how* prognostic research is performed with guidance for budding prognosticians on how to improve their research. Systematic reviews can also be considered fundamental as they provide a meta-analysis of how prognostic research is currently being performed with praise and criticism where approriate.

Stratified medicine aims to provide patients with a focused and specialised decision on what treatments they should receive which will produce the best results for that patient (or lowest cost for the healthcare provider)[4]. Medical textbooks provide information on the prognosis of diseases in general, but the goal of stratified medicine is to tailor the treatment to an individual patient rather than treating everyone with a disease the same[40]. The idea is to provide the best treatment for the individual which is not necessarily the best average treatment for the disease (i.e. a specific patient might be predicted to respond better to a treatment than the average patient)[40]. It is widely believed that as we move into a more data driven society, we will be able to use this data to group patients into smaller and smaller strata for better and more tailored treatment[41,42].

3.1 Prognostic Factor Research

The aim of prognostic factor research (the second stage of prognostic research) is to discover which factors can have an effect on the prognosis of a patient with a specific condition and is a subclassification of predictive factor research. Predictive factor research is the starting point to producing a usable predictive model as without knowledge of which variables *could* affect an outcome, it is impossible to figure out which variables *will* affect an outcome. For example, in the development of the NPI, it was known that cancer tumour grade is a prognostic factor in cancer patients as a higher grade correlates with higher mortality[5]. Note the use of the word correlate in the previous sentence. A recurring theme of statistics is the idea that correlation does not imply causation. Margarine consumption is correlated with divorce rates, Nicolas Cage films are correlated with pool drownings and chicken consumption is correlated with oil imports[43]. Predictive (and therefore prognostic) factor research is not aimed at discovering causal relationships, but merely uncovering correlations[40]. These factors may indeed be causal, but this is not a requirement and it would be up to an aetiologist to prove a causal relation and not a prognostician or diagnostician[2]. In neonatal care, skin colour in the Apgar score is a predictive (diagnostic) factor but not a causal one[40].

Counter to the idea that prognostic factors aren't always causal, they are *always* confounding factors for the event they predict. Thus prognostic factors should be taken into account when planning clinical trials as if they are misbalanced across the arms (or not accounted for in some other manner), they can cause biases in results[2] and sometimes adjusting clinical trial results by a prognostic factor can affect the interpretation of the results[44]. If a prognostic factor is causal, then by directly affecting the prognostic 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 in treatment of the outcome.

Some predictive factors might only elicit a response for patients on treatment (such as a metabolic effect removing a drug from the patient's system faster) and give no response to those without treatment[4]. For example CYP2C9 and VKORC1 genotypes have an influence on patients being treated with warfarin, but have no effect on patient's risk of stroke without the treatment[45]. This is called a differential response. A positive result from a *BRCA2* test can be an indicator of a risk of breast cancer amongst women[46]. If two women, one young and one old, both test positive for the gene, the young woman would likely be advised to have a mastectomy, whereas the older woman would not. This is because the younger woman is of higher risk of dying from breast cancer (even though breast cancer has not been diagnosed) than the older woman. A prospective study by Fliser at al[47] found that Fibroblast Growth Factor 23 (FGF23) plasma concentrations was a prognostic factor for the progression of CKD. Haemoglobin A_{1c} levels should be routinely measured in patients with diabetes as it is a prognostic factor for vascular events[48]. Systemic Lupus International Collaborating Clinics (SLICC)/American College of Rheumatology (ACR) Damage Index (SDI) is used to measure the level of permanent damage in a patient with Systemic Lupus Erythmatosus (SLE)[49] and is thus a prognostic factor for cardiovascular related deaths[50]. Different ethnic groups can have vastly different prognosis in regards to cardiovascular disease[51,52] and so ethnicity (and family history) are prognostic factors for most clinical events. These are all examples of prognostic factors from a wide variety of clinical sources.

It is unfortunate, however, that Riley et al[53] 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[2].

3.2 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[3] and they are designed to augment the job of a clinician and are not intended to replace them outright[40]. Diagnostic prediction models can be used to indicate whether a patient is likely to need further testing to establish the presence of a disease[37,38]. Prognostic prediction models can be used to decide on further treatment for that patient, whether as a member of a certain risk group, or under a stratified medicine approach[37,38]. 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[40]. 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[18]. 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 clinically useful model is not as simple as just using some available data to develop a model,

despite what a lot of researchers seem to believe. To quote Steyerberg et al[3] "To be useful for clinicians, a prognostic model needs to provide validated and accurate predictions and to improve patient outcomes and cost-effectiveness of care". This means that, although a model might appear to be useful, its effectiveness is only relevant to the population it was developed in. If your population is different, then the model will behave differently. Bleeker[54] developed a model to predict bacterial infections in febrile children with an unknown source. The model scored well when assessed for predictive value in the development data set, 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.

3.2.1 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, including MSMs.

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[2]. 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[5], or ones which could have a more general effect on overall health, such as socioeconomic and ethnicity variables[55]. 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[56]. 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[7].

When developing a predictive model, the size of the dataset being used in an important consideration. There should be at least 10 events for every potential predictor[57,58], this value is known as the Events Per Variable (EPV). If there aren't enough events to satisfy this criteria, then some potential predictors should be eliminated (for example, using clinical knowledge) before any statistical analysis is performed on the data[59]. This will help to minimise selection bias and model instability. In general, it is also recommended that this development dataset contain at least 100 events (regardless of number of potential predictors)[39,60,61]. A systematic review by Counsell et al[62] found that out of eighty-three prognostic models for acute stroke, less than 50% of them had more than 10 EPV. Having a low EPV can lead to overfitting of the model which is a concern associated with having a small data set. Overfitting leads to a worse prediction when the model is used on a new population which essentially makes the model useless[34]. 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[39]. 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[39]. 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[39]. 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 similar, 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[35,63]. Derivation of the QRISK2 Score[7] (known later as QRISK2-2008)[10] randomly assigned two thirds of practices to the derivation dataset and the remainder to the validation dataset. The NPI model was trained on the first 500 patients admitted to Nottingham City Hospital after the study began[5] and later validated on the next 320 patients to be admitted[64], this validation was not performed at the same time as the initial development and is thus an external validation.

If a sufficient amount of data is available and it has been taken from multiple sources (practices, clinics or studies), then it should be clustered to account for heterogeneity across sources[65]. It is important that any sources of potential variability are identified (such as heterogeneity between centres) as this can have an impact on the results of any analysis[1,39]. Heterogeneity is particularly high when using multiple countries as a source of data[66] or if a potential predictor is of a subjective nature, which leads to discrepancies between assessors[67]. Overlooking of this clustering can lead to incorrect inferences[65]. 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[40].

During development of any model, using only the complete case patients immediately reduces the power of the results due to the lower number of patients in the population, but it also can introduce bias from patients not being missing at random[34]. Multiple imputation can be used to fill in the gaps in missing data and it is recommended (over complete case or single imputation) during development to avoid this bias[68]. Complete case analysis is a viable option only when a small percentage of patients ($< 5\%$) do not have all available data[69]. During development of the NPI, only 387 of the 500 patients were included in the study due to missing data in the other 113[5], however QRISK2 did use multiple imputation methods to deal with the missing data[7]. In a review of CKD predictive models by Collins et al[12], out of eleven models, four conducted complete case analysis only and only two conducted multiple imputations on the missing data. The remaining five studies did not mention missing data at all. Small data sets, inappropriate handling of missing data and lack of validation are common issues in prediction model development[37,38].

A prediction model researcher needs to select clinically relevant potential predictors for use in the development of the model[34]. 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[37,38]. Potential predictors with high levels of missingness should be excluded as this missingness can introduce bias[34]. One key fact that many experts agree on is that categorisation of continuous predictors should be avoided[3,12,34,44,59,70] as it retains much more predictive information to keep them as continuous[44]. The cut-points of these categorisations lead to artificial jumps in the outcome risk[59]. It is also worth noting that cut-points are often either arbitrarily decided or data-driven, the latter leading to overfittingl[59]. If categorisation is performed, clear rationale should be provided with an acknowledgment that this will reduce performance[12,70]. When applying a model to a new population, extrapolation of a model should also be avoided[71] and so to aid in this, the ranges of continuous variable should be reported[12]. This is especially true for age. QRISK2 was derived in a population ranging from 35 to 74 years of age and so can not accurately be applied to patients out of this range[7]. This range was later extended with the updated versions[8] and currently can be applied to patients aged 25-84[9].

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[59]. 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 increase 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[59]. If possible, subject matter knowledge should guide the decision, but obviously this is not always available[59]. 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[34]. Univariable analysis does not account for any dependencies between potential predictors and so any correlation that exists between them can cause a bias in the results. Despite its clear weaknesses, many prognostic studies still use univariable analysis to build their models[72].

The NPI predictive model includes lymph-node stage, tumour size and pathological grade to identify patients with a poor prognosis with much better discrimination than would be possible if only one of these factors were used in isolation[5]. The develpoment of the model began with nine potential predictors, of which three were considered to be statistically significant ($Z > 1.96$, $p < 0.05$) in a Cox model[73] and so were included in the final model which was simplified to $I = 0.2 \times \text{size (in cm)} + \text{stage} + \text{grade}$.

Backwards Elimination (BE) involves starting with all potential predictors in the whatever kind of model we are using and removing ones which do not reach a certain level of statistical significance (for example, 5%) one at a time until all remaining variables are significant. Forward selection begins with no variables and adds one at a time based on similar criteria. Under either of these methods, a lower significance level will exclude more variables[34]. Backward elimination of variables is preferable over forward selection[74]. However, it has been argued thoroughly and convincingly by Sauerbrei et al[59] that a better choice for predictor selection (without subject matter knowledge) is using the Multivariable Fractional Polynomial (MFP) method. This method increases the number of potential predictors by applying different combinations of fractional powers from a predefined subset to the continuous potential predictors[59]. This essentially expands the list of potential predictors to include these fractional powers and then performs BE on this larger predictor set of variables. If we can be convinced that all variables are linear (or linear under a feasible mapping), a simple BE technique is just as good of a choice as MFP.

A variant of the conventional BE technique involves using the Akaike Information Criterion (AIC) rather than statistical significance. For this method, models are generated by removing each remaining potential predictor

individually from the pool and calculating the AIC of each of these models as well as the AIC for the full model. If we have m potential predictors then this produces $m + 1$ values for the AIC. The model with the lowest AIC is the best of these models and so we repeat the process with the pool of predictors used for that model. This is equivalent to the conventional BE method using 15.7% as the significance level.[75].

QRISK2 checked for non-linearity amongst continuous potential predictors using fractional polynomials[7,76] as well as certain interaction terms (in particular interactions with age). Where an interaction in factors is identified in a study, this can be a useful indicator of a differential response and should be investigated further[4]. If a predictor is expensive or invasive to measure, it might be better to include a less significant predictor which is easier to come by[40]. A limiting factor for some prognostic models is that the prognostic factors they measure are not readily available or are not used in routine care[3].

Once developed a prognostic models can be used to create risk groups for a population. Risk groups should be defined by clinical knowledge rather than statistical criteria[35]. Grouping patients into risk groups is not as accurate as using the specific model to provide an estimated risk[3]. The original development paper for NPI, patients were classified into three risk groups, Low ($I < 3.4$), Medium ($3.4 < I < 5.4$) and High ($I > 5.4$)[5]. The followup paper extended these groups to be: Very Good ($I \leq 3$), Good ($3 < I \leq 4$), Moderate ($4 < I \leq 5$), Poor ($5 < I \leq 6$) and Very Poor ($I > 6$) with annual percentage mortality rates of 1.5, 3.5, 6, 20 and 32 respectively[64].

Since QRISK2 has been developed in association with Egton Medical Information System (EMIS), there is the ability to automatically apply a QRISK2 measurement to all patients in EMIS and thus produce a list ordered by risk score to establish which patients are most likely to suffer a cardiovascular event in the next 10-years and to which patients resources can be prioritised[7]. Because of this, during development, it was estimated that QRISK2[7] would be generalisable to around 80% of practices in the UK although they acknowledged that the model should still undergo external validation. Derivation of QRISK2[7] was not done on patients with a history of cardiovascular disease and so it *cannot* be applied to them. It was acknowledged in the original QRISK2 development paper[7] that it would require updating, which now happens annually with more up-to-date data[9].

The systematic review by Counsell et al[62] found that only four of the eighty-three models met their 8 key quality criteria. Models were assessed on external validity, internal validity, evaluation and practicality. As mentioned previously, one of these conditions was to have an EPV greater than 10, and the other 7 were: adequate inception cohort, less than 10% loss to followup, prospective data collection, valid and reliable outcome, age as a candidate predictor, severity of condition as a candidate predictor, use of stepwise regression. None of the four models which satisfied the criteria had been externally validated. Over 150 different predictors were considered across the review with most of them only occurring in only 1 or 2 of the modes. Of the 18 variables that were assessed in at least 5 models, only 3 were significant in more than 80% of the relevant models. This demonstrates a lack of cohesion between research groups with a disregard for previously developed models and a preference for researchers to use available data sources to develop new models rather than update existing ones[3,35].

In their systematic review, Collins et al[12] assessed ten study papers. In this review, they identified 97 potential predictors that were considered across the ten studies, with 58 of them only being considered in a single study with a median of 4 potential predictors per study and only six of the studies justified their choice of potential predictors. Only three of these studies reported the age range of the patients involved and only two studies assessed linearity in the predictors. A review of 47 papers by Mallett at al[77] found that the reporting of prediction model development in cancer was very poor in all measures. Another review of 71 papers by Bouwmeester et al[78] found that even the reporting of prediction model development in high-impact medical journals was incredibly poor.

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

A very small number of developed models are currently routinely used in clinical practice[12], however their use is becoming more common, with more and more healthcare providers recommending their use[37,38]. This current lack of clinical implementation may be due to the use of inadequate reporting techniques of predictive models can cause problems when evaluating the potential usefulness of the model[12]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[37,38]was developed in an attempt to combat this inconsistency. It is described as "a guideline specifically designed for the reporting of studies developing or validating a multivariable prediction model, whether for diagnostic or prognostic purposes"[37,38]. As acknowledged in the QRISK2 development paper[7], as new technologies arise (such as eHealth Recordss (EHRs) and genomic measurements), it is important to update existing models with added information. Models should be constantly updated and re-validated as populations and ambient health effects change over time[3].

3.2.2 Model Validation

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

The external validation of a prognostic model is considered by some to be more important than its development as it demonstrates the generalisability of the model, without which, a prediction model is essentially useless[3]. Because of this, the TRIPOD guidelines strongly recommends researchers to perform an external validation on their models, whether as part of the initial development paper or a subsequent one[37,38]. This also means that unvalidated models should not be used in clinical practice[35,54]. Despite this, they are often accepted as they are without being rigourously tested[54]. This means that clinicians should be wary when using predictions from models that are yet to be externally validated[40], however, even models which perform moderately under external validation are likely to be better than a clinician's assessment[63]. Unfortunately, external validation studies are scarce, especially when compared to development studies[35,80]. Hopefully, the ability to access big datasets such as EHR or Individual Participant Data (IPD) will allow external validation studies to blossom in the coming years[39]. For example, the QRESEARCH database, which was used to create the QRISK and QKIDNEY scores database contains 24 million patients from 1300 general practices[81]. The database contains longitudinal, demographic and mortality data on the individual patient level as well as many other factors that can be used for clinical prediction[7].

Discrimination is the ability of the model to separate out patients who are more likely to have an event from those that are less likely[35,82,83]. The D-statistic is a common measure of discrimination for time-to-event data. To find the D-statistic, we split the validation dataset into two at the median value for the prognostic index. The D-statistic is then the log hazard ratio between these two groups[39]. A higher D-statistic indicate greater discrimination[39,84]. The c-statistic is the probability that if we choose two patients at random, the one with the lower risk score will have the event and the one with the higher score will not[39]. For binary outcomes, it is known as the area under the Receiver Operating Characteristic (ROC)[39,54]. Higher values of the c-statistics implies better discrimination[7]. The c-statistics is between 0 and 1 with 1 indicating perfect discrimination and 0.5 indicating no better than chance[39]. If a c-statistic is between 0 and 0.5, the results from the model can be reversed to provide a model with a $1 - c$ c-statistic (e.g. if a model has a c-statistics of 0.25, by reversing the outcome, you create a model with a theoretical c-statistic of 0.75, however this new model would have to be externally validated). The c-statistic has been criticised for its inability to detect meaningful differences[85] and is commonly between 0.6 and 0.85 for prognostic models[36]. The Framingham Score, which is used to predict risk of cardiovascular event, was found to have a c-statistic of around 0.70[86]. Fraccaro et al[87] conducted a study using a large EHR dataset ($n = 178,399$) to validate seven models that predict the onset of CKD and found that all seven had a c-statistics of around 0.9 indicating high levels of discrimination.

Calibration is the ability of a model to accurately predict the number of events in a group of patients (e.g. among patients who are predicted a 10% chance of having an event, 10% of them should have the event)[35,82,83]. It can be assessed graphically by grouping patients into deciles by their predicted (or expected) risk and plotting the average expected risk against the observed outcome (i.e. the percentage of patients in that grouping who had the event)[35,39]. A line-of-best-fit can be added to this plot which demonstrates the calibration of the model. The slope of this line (known as the calibration slope) should ideally be 1[39]. If it is less than 1, some prediction are too extreme, whereas if it is more than 1, the predictions are too narrow. The calibration slope of all three versions of QRISK2 ranged between 0.92-0.95 for men and women which indicates a very good calibration[10]. Calibration plots can be accompanied by a Hosmer-Lemeshow test[88]. This test is similar to a χ^2 test where patients are grouped by predicted risk. Another measure of calibration is the Expected/Observed number of events (E/O) which is simply the number of expected events across the population (i.e. the sum of the predicted probability that each patient will have an event) divided by the observed number of events. Again, the E/O should ideally be 1[39]. If it is less than 1, the model is under-predicting if it is greater than 1, it is over-predicting. It should be borne in mind that summarising validation statistics is not always adequate and that assessing the validation of a model on different subgroups is more beneficial[39]. For example, the Framingham Score has an E/O of 1.03 over an entire population[10]. However, in women aged between 40 to 64, it over-predicts and in women aged 70-74, it under-predicts.

Internal validation will provide an indication towards whether there is an inherent lack of generalisability of the model[3]. Three common methods for perfoming internal validation are sample splitting, bootstrapping and cross-validation. Cross-validation involves randomly splitting the dataset into m subsets (e.g. in 10-fold cross-

validation, you have 10 subsets) and then developing a model on $m-1$ sets and then validating it on the other set. This is repeated m times for each data set giving m new models which can be compared to the model produced from the entire dataset and validation data for each of these models. When bootstrapping, if we have a population of size n, then we randomly select (with replacement) n patients from the population and validate on this new population. This is repeated a prespecified number of times (e.g. 1000 times) to give validation measures. The final method used for validation is sample-splitting where model developers randomly split their population into two subsets, a training set and a testing set, usually in a 2:1 ratio. The training set is used to develop the actual model which can then be validated against the testing set. Both cross-validation and bootstrapping provide sets of values for each of the validation measures we are using which can be aggregated to provide estimates (e.h. means and confidence intervals) of what these values would be if the model were validated externally. These estimates would only be relevant to an external population which is extremely similar to the development dataset.

Researchers can assess the levels of overfitting, optimism and miscalibration using an internal validation[37,38]. Model optimism can be estimated by taking the difference between the performance of the model in the bootstrapped datasets and the performance in the original development set[54]. Bootstrapping techniques are consdered to be an effective and efficient method for internally validating a dataset against the current population (as opposed to cross-validation or split-sample methods)[54]. Bootstrapping provides a nearly unbiased estimate of the effect of the predictive accuracy of a model[82,89], it can provide a shrinkage factor to partially calibrate a model to improve performance[54]. However, bootstrapping does not change the underlying population and so is not quite thorough enough[90] and so external validation is still more important to the usability of the model than internal validation. The ability to translate a model into a new population is called its generalisability. Because it uses the development data set, internal validation provides no information on the generalisability of a model, but merely indicates how well the model is calibrated for and discriminates in the development population[34,35]. Although poor discrimination and calibration can imply poor generalisability, good discrimination and calibration does not imply good generalisability[54].

Prediction models are often overfitted, meaning they perform better on the data for which they have been developed than on external data[54]. Different settings, both geographic and temporal can cause poor performance in external validation[35,36], for example, models developed in primary care are different from those developed in secondary care due to the higher rate of more severe conditions in secondary care[91]. When applying a model to a new dataset, it is important to compare the populations of the derivation dataset and the new dataset to ensure compatibility of the model[54,92,93]. This is done through a comparison of the case mix of a population. The case mix is the defined as the distribution of all of the potential predictor variables and the outcome variables (incuding those which do not make it into the final model)[36]. If these are similar between two populations, then we can assume the populations as a whole are relatively similar. Models will be more generalisable to a new population if the case mixes of the two populations are similar[36]. Applying a prediction model to a similar population can improve the chances of the model fitting the dataset well, but this will reduce the applicability of the model to other populations in the future and might cause unwarranted confidence in the predictive abilities of the model. High heterogeneity in the development population is an advantage as low heterogeneity leads to low discrimination (if patients are similar, their prognosis would also be similar) which can therefore reduce the generalisablility of the model[39]. In some cases the difference between the development population and the validation population can be enough to render a model useless after external validation[54]. It would be difficult to ascertain how often this happens as external validation studies which results in poor performance are rarely published. If performance is not consistent across populations then users should be made aware of this and might even be advised to use a different model in certain cases[39]. Validating models developed in one setting in another setting is useful, but is expected to produce less than ideal validation statistics (discrimination and calibration)[39].

If, during external validation, a model does not reach a high enough standard of calibration, but has good discriminative ability, it is more useful to recalibrate the model (by introducing a multiplicative factor) than to develop an entirely new model. The recalibrated model may need to be revalidated in another population[36]. Recalibration involves adjusting the coefficients of the model to achieve a calibration slope closer to 1[36]. If recalibration is not possible, updating the model in other ways is preferable over creating a new one and there are many techniques which can combine the original model with the new dataset[94]. However a model is updated, the new model should always be re-assessed for external validity[36]. Model development often takes precedence over model validation as a primary goal of research and most models don't make it out of the development stage[3,39]. It would be more productive for the field of prognostic models (and the specific fields for which the models are relevant) if there was a concerted effort to validate models more often rather than constantly developing new ones addressing the same questions[3]. Unfortunately, external validation studies will often be abandoned if poor performance is found, an idea that perforates the entirety of science through the hesitation to publish insignificant findings and a lack of replication studies[95]. Once an external validation study has been abandoned, researchers will then develop an entirely new model from their datasets and not publish the poor validation results[36]. This means that rather than having a developed and validated model (which was the original intention of the project),

we now have two unvalidated models for the same (or similar) outcome. This might cause another team to try and validate the original model (or the new model), leading to the same issue ad infinitum.

During the original development, Hippisley-Cox et al used sample splitting techniques for the internal validation[7] and compared the performance of QRISK2 with the modified Framingham Score[96]. For this internal validation, patients were deemed as "High Risk" if they were predicted a 10-year risk of cardiovascular event of $> 20\%$ for each score seperately. They then compared patients who were considered "High Risk" under one measure and not under the other to determine how a change in model usage (e.g. from Framingham to QRISK2) would impact patients. Overall, of the patients in the QRISK2 High Risk set, 23.3% of them had an event over the 10 year observation period and in the Framingham High Risk set, 16.6% had an event, indicating an underestimation by the modified Framingham compared to an overestimation by QRISK2. Amongst the two groups of reclassified patients, those in the QRISK2 High Risk set had an annual incidence rate of 30.6% and 35.2% for men and women, respectively. This is in contrast to the Framingham High risk set which had 25.7% and 26.4%. This indicates that, at the 20% threshold, QRISK2 identified a more at risk population than Framingham did.

In 2012, Collins et al[10] externally validated three versions of QRISK2 (from 2008, 2010 and 2011)[7,8]. They used the The Health Improvement Network (THIN) cohort and each patient was given a predicted risk score based on the three QRISK2 equations as well as a fourth score based on a the modification of the Framingham score[96–98]. The regular Framingham Score calculates a risk for coronary heart attack and a risk for stroke separately, the modified score sums these values together and applies specific multiplicative effects depending on the patient (1.4 for south Asian men and 1.5 for a family history of coronary heart disease). Since the two risk scores are not necessarily independent this can result in a higher than 100% risk for some patients (especially when this is multiplied by 2.1 if a patient is south Asian with a family history of coronary heart diseases). When this analysis was performed, QRISK2-2011 was the current version of QRISK, it has now been updated to QRISK2-2016[9] and is expected to be updated to QRISK2-2017 this year. When the QRISK2 model was developed, the National Institute for Health and Clinical Excellence (NICE) recomendations were to use the modified Framingham Score to assess patient's coronary risk, however by the time the external validation was done, NICE recomended that physicians choose between Framingham and QRISK2 when performing this assessment based on their own experience with the two models[10].

Multiple imputation was used by Collins et al[10] to deal with missing data for the THIN cohort[99]. Amongst patients in the THIN cohort, the modified Framingham score over-predicted for most patients giving a very shallow calibration slope[10]. Most patient's QRISK2-2011 scores were very similar to their QRISK2-2008 and QRISK2-2010 scores with almost all of them having their updated scores being within 3% of the older versions[10]. However in the internal validation, the calibration and discrimination were only summarised across all 176 practices which ignored potential heterogeneity in the population from practice to practice[7,39]. A similar thing occurred in the external validation which ignored between-practice heterogeneity ($I^2 = 80.9\%$)[10,39]. The external validation paper reported a 95% confidence interval for the c-statistic of 0.826 to 0.833 using data across 364 practices. However due to this between-practice heterogeneity, if the process were repeated with a new practice, we would predict it having a 95% confidence interval of 0.76 to 0.88. The conclusion of the external validation paper was for NICE to recommend that healthcare professionals abandon the Framingham score in favour of the QRISK2-2011 model, or at the very least to use a recalibrated version of the Framingham score[10]. Due to the improved predictive ability of the QRISK and QRISK2 scores and the results by Collins et al, the recommendation by NICE to use the modified Framingham Score was withdrawn in 2014 and clinicians are now advised to use the most recent version of the QRISK2 score available[9,10,97].

For the validation study of NPI[64], it was assessed prospectively in a group of 320 patients at the same hospital as used for the development[5]. All 707 patients (387 original and 320 new) in the NPI validation study[64] were assessed by the same pathologist and under the care of the same surgeon as the development study[5]. As well as the three factors in the NPI model[5], the validation study[64] also collected data on menopause status and Oestrogen Receptor (ER) as these were close to significant in the original study ($Z > 1.5$, $p < 0.134$). It is demonstrated graphically in the NPI validation study[64] that the risk groups produced in the original study[5] are viable in the new population. The Cox analysis[73] was also re-run on the original 387[5] patients as more followup time was now available and the resulting coefficients were similar to the original study indicating the original model performs well in the extended dataset. This allowed Todd et al[64] to update the NPI risk groups[5] to include five groups rather than the original three which allows for better stratification as mentioned earlier.

Fraccaro et al[87] performed an external validation of multiple models designed to predict the onset of CKD. It was described as "The first comprehensive head-to-head comparison study of multiple CKD prediction models on a large independent population"[87]. Five of the models assessed needed to be recalibrated to suit the population of Salford, UK. QKidney[100], which was developed in the UK performed the best. The model developed by Bleeker[54] to predict cause in febrile children failed during external validation. The ineffectiveness of Bleeker's model[54] was confirmed by refitting a multivariable model to the new dataset which provided significantly different

estimates for the coefficients. This was, however, useful as some of the predictors were still considered to have an effect on the outcome prediction, just at different magnitudes from the original, which implies that they can be useful as potential predictors in future models. Bleekers results stand as a testament to the necessity of external validation[54].

During development, models should be reported with transparency to allow for accurate recreation during validation[12]. If the model fails during external validation, then this transparency will allow researchers to establish *why* it failed and to ensure the same mistakes are not repeated. This transparency should therefore be retained when validating a model and if it performs poorly, this should still be reported. For example, if a model is found to perform poorly, and the researchers develop a new model, this should be included in the rationale for the development[37,38]. When an external validation shows that a model is ineffective, this may be down purely to "bad luck" in two populations, although there would have to be further external validation to prove this conjecture which would be unlikely to happen as if a model has failed external validation, it is usually discarded without further investigation. If the model has been developed on the back of a failed validation, then when validating this new model, researchers can also re-validate the original model to check whether it was "bad luck" in the model. Models should also be re-evaluated frequently to ensure they are kept as up-to-date as possible[7,35,36]

3.2.3 Model Impact Evaluation

As predictive models are considered to be health technologies, once they been developed and validated with good discrimination and calibration (or been re-calibrated), their usefulness in the real world will need to be assessed[3]. This is done through an impact study. Although models with near perfect discrimination and calibration may not need an impact study to evaluate their usefulness[36]. From a health economics standpoint, it is useless to implement a model which costs more for the healthcare service than the current standard unless the improvements to population health are substantial. Impact studies should *always* be performed before a model gets used in clinical practice as they assess a models robustness and generalisability more thoroughly than an external validation[39,54].

The impact of a prognostic model can be assessed in a manner similar to that of an RCT where some patients are provided treatment based on decisions made from a prognostic model and the control arm receives the best alternative treatment[3,36]. Impact studies can be performed as a before and after comparison which may be simpler to implement, but can be sensitive to ambient changes in the clinical environment[101,102]. Randomisation in an impact trial can happen at the patient, doctor or centre level. Higher levels (centres) is much preferable over lower ones (patients) as this lowers the likelihood of cross contamination of the intervention (such as a single doctor treating patients on both arms of a study, or two doctors in the same centre discussing results)[103]. The intervention of an impact study can be an assistive or a decisive approach[36]. Assistive approaches provide the clinician with the patient's risk score as indicated by the model and allows the clinician to use his or her own judgment on how to proceed with treatment. Decisive approaches explicitly tells the clinician what decision should be made based on the model (e.g. whether or not to prescribe statins). The STarTBack trial[104] was an impact study which compared the use of stratified care with traditional "best care" on patients with lower back pain. The results showed that the stratified care reduced the risk of disability and lowered the cost of care compared to the control arm. The NPI is widely cited and used, however it's impact has hardly been assessed[3,5].

4 Competing Risks and Multi-State Models

Many diseases are measured in stages of progression or as types or variants. Often, patients can switch from one of these stages to another whilst they are being studied. If being in a different stage of the disease is believed to affect the way that the patient's condition behaves then it is important to account for this when modeling a disease. The simplest way is to have the disease stage/type as a covariate and ensure that it is updated accurately. However, if it is believed that a disease behaves wildly differently when at different stages, then this might not be feasible, especially if the stage can interact with other covariates. The solution to this is to use MSMs to map patients progression through the different stages of the disease[17], where each stage is modelled as a state in the MSM.

From survival analysis, a hazard function is a measure of the intensity of moving from one state to another (whether that is from alive to death, functioning to non-functioning or something more complicated as in an MSM). If we have T be the random variable defining the time of the event (or transition), then a hazard function is usually defined as[105,106]

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

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

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

We can simplify this equation to $S(t) = \exp(-\Lambda(t))$ if we define the cumulative hazard function to be

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

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

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

A function that is useful in demonstrating the survival of a population is the Kaplan-Meier estimate which is defined as:

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

where n is the number of patients and patients $j = 1...n$ ordered by t_j, the time of event and d_j and n_j are the number of patients having an event at t_j and the number of patient still at risk (i.e. in the risk set) at time t_j. Kaplan-Meier estimates assume independence between the event we are modeling and censoring[106]. See figure 2 for a typical Kaplan-Meier plot for two populations.

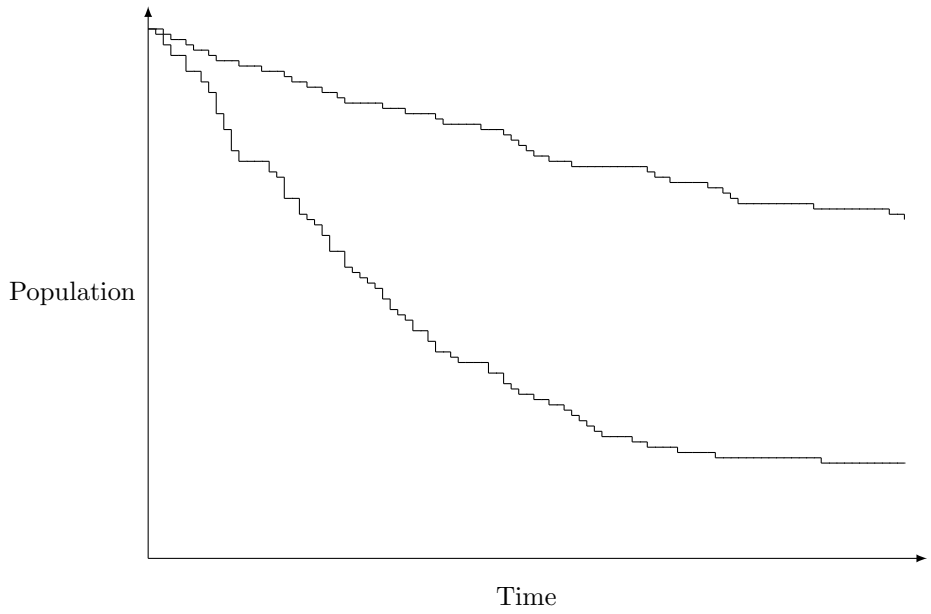

Figure 2: Plot of a Kaplan-Meier estimators for two populations

As mentioned in Section 3.2.1, there are many different kinds of models that we can use data to produce. Most models can be rearranged to be a linear relationship between predictors and outcome. A linear relationship being one of the form:

$$Y = \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_m Z_m = \beta^T Z$$

where the βs are the coefficients found from the data and the Zs are the covariates or predictors. In survival analysis, the two most common models are logistic regression and Cox proportional hazards method[107]. For binary outcomes, a logistic model uses the data to create a linear relationship between in the predictors and the log odds of the outcome, known as the logit, which can then be converted into a probability:

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

$$p = \frac{e^{\beta^T Z}}{1 + e^{\beta^T Z}}$$

where p is the probability of the binary outcome, whether this is a positive diagnosis or an event happening within a certain time. A Cox proportional hazards model assumes a baseline hazard function which is affected by constant

multiplicative factors depending on the covariates. These multiplicative factors are written as the exponentials of the linear coefficients and are unchanging over time. The Cox model can be written as:

$$\lambda(t|X) = \lambda_0(t) \exp\left(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_m Z_m\right) = \lambda_0(t) e^{\beta^T Z}$$

To provide estimates of probability of an eent for the logistic model, we simply plug in the relevent predictors. However, for the Cox model, we also need to estimate the baseline hazard function (usually based on the median values of the population[7]).

When data is assessed under a Cox model, the exponentials of the coefficients, $\exp(\beta_j Z_j)$ are usually refered to as Hazard Ratios (HRs). HRs are useful in a lot of different areas of medicine as they provide a definitive multiplicative value for the effect that a certain covariate has on the outcome. For example, if the HR of a binary predictor is 2, then a patient with that binary predictor will be twice as likely to have an event on any given day than a patient without the predictor. HR provide a relative risk and so they are used a lot in RCTs as we can say that a treatment provides a specific multiplicative effectr compared to the control arm.

4.1 Competing Risks

A CR can be thought of simply as survival analysis where a cause of death, $D \in \{1, ..., k\}$ is also observed[108]. When patients are recovering from a disease, more than one event can play a role, but often one event is of more interest than the other[106]. This competing event can also prevent the event-of-interest from occurring. For example, if we are modeling discharge from hospital after surgery (event-of-interest), patients can also die whilst in hospital (competing event) which prevents the former from happening. Depending on clinical context, non-administrative right censoring can be modeled as a competing risk[108] because censoring times are not always independent of event times[106]. If healthier patients are less likely to use medical services, then they are more likely to be lost to followup meaning a negative correlation with event time and vice versa if a less healthy person is more likely to leave a study (e.g. they become too ill to continue in the study). If the competing event and event-of-interest are not independent, then this can cause bias in the Kaplan-Meier estimator[106]. Issues can arise with the naive Kaplain-Meier approach in CR, wherein the probability of having the events sum to more than 100%, even though the events can not occur together[106].

We can define the cause-specific hazard function for an event h as

$$\lambda_h(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}\left(T \leq t + \Delta t, D = h | T \geq t\right)}{\Delta t}$$

where D is the type of event (e.g. the cause of death). Cause-Specific Hazard (CSH) estimates may be found by treating the data as simple survival data and the competing events as a censoring event[108]. However, this may not be suitable if there are coefficients in common between the transitions (e.g. every covariate has a different effect on each hazard)[108].

In CRs, the survival function is the probability that none of the events have happened and so we define:

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

which is also know as the marginal survival probability, and the probabilities of transitioning to a particular state as:

$$\mathcal{P}_h(t) = \int_0^t S(u-)\lambda_h(u)\, \mathrm{d}u, \qquad h = 1, ..., k$$

Since $S(u-)$ is the probability of surviving up to a time, u, and $\lambda_h(u)$ is the instantaneous hazard of having an event, $S(u-)\lambda_h(u)$ is the probability of surviving until u and then having the event. Since this is an instantaneous occurance, we have to integrate between 0 and t. $\mathcal{P}$ is usually referred to as the Cumulative Incidence Function (CIF). The CIF for a transition depends on all the other transition intensities, through the $S(u-)$ in the integration[108]. If we let T_i be the survival time and D_i be the cause of death for patients $i = 1, ..., n$, we can express the likelihood function of the hazards[108] as:

$$L = \prod_{i=1}^{n} S(T_i) \prod_{h=1}^{k} \lambda_h(T_i)^{I(D_i = h)}$$

which can be used to estimate the λs.

The underlying issue surrounding CR is the assumption that if one cause of failure is removed, the others would behave exactly the same[109,110]. Fine and Gray[111] developed the Proportional Subdistribution Hazards method (PSHM), which is a method for calculating the cause specific HRs using the CIF[106], $\mathcal{P}_h$, and based on the regular Cox model[107]:

$$\widetilde{\lambda}_h(t) = \widetilde{\lambda}_{h0}(t) \exp\left(\beta_h^T Z\right) = \frac{\partial}{\partial t} \log\left(1 - \mathcal{P}_h(0,t)\right)$$

Subdistributions get that name from the fact that they are always strictly less than 1 and so aren't "proper" distributions[106]. In the subdistribution hazard[111], patients who suffer from another event (other than event k) remain in the risk set, whereas for the CSH, they do not. Both the CSH and PSHM can easily be extended to include clustering[112].

An example of using a CR model in the real world involves mortality after an Acute Myocardial Infarction (AMI)[108]. Patients could either have a sudden Cardiovascular Disease (CVD) related death, a non-sudden CVD related death or a non-CVD related death (i.e. death from other causes). Using age and gender as covariates for a simple CSH Cox model[107], HRs can be calculated for each cause of death, see Table 1.

	Cause of death			
	Non-CVD	Sudden CVD	Non-sudden CVD	All causes
Age per 10 years	2.06 (1.84 - 2.29)	1.56 (1.43 - 1.70)	2.13 (1.96 - 2.31)	1.90 (1.80 - 2.00)
Gender = Male	1.24 (1.00 - 1.53)	1.34 (1.11 - 1.63)	1.05 (0.90 - 1.23)	1.18 (1.06 - 1.31)

Table 1: Hazard ratios for the covariates for the different causes of death taken from Andersen et al, 2002[108]

Having a more prognostic attitude towards medicine, as opposed to a diagnostic one, could help to reduce wasted expenditure under a CR scenario. For example, a patient with high blood pressure automatically has a lower risk of death from prostate cancer (because they are more likely to die from prostate cancer) than a similar patient with normal blood pressure and so the first patient would gain less from a prostatectomy[46]. Diagnosis often dichotomises a patient's condition into having or not having a particular condition or disease[46]. Oftentimes, these conditions are actually a continuous spectrum with an arbitrary cut-off point for diagnosis. This is particularly true in conditions of high prevalence amongst industrialised nations, such as Type II diabetes (having glucose level >125 mg/dL)[46].

4.2 Multi-State Models

Just as CRs are an extension of survival analysis, MSMs are an extention of CRs[106]. CRs are a subset of MSMs where there is only a single initial/transient state and multiple absorbing states which a patient can transition into[105,108]. An absorbing state is a state which has no transitions coming out of it (such as death or discharge), a state which is not absorbing is called transient[105]. MSM can be used to model many different conditions if we believe that the different states of the condition have an inconsistent effect on the outcome. By this we mean that being in different states can cause the hazard function to behave very differently. In this case, we simply use different hazard functions depending on the state and therefore, we have an MSM.

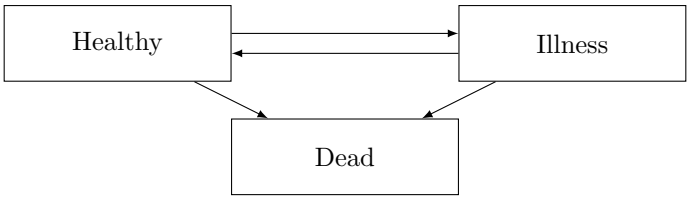

Figure 3: A Reversible Illness-Death Multi-state Model

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

$$\mathcal{P}_{hj}(s,t) = \text{Prob}\left(T(t) = j | T(s) = h, \mathcal{H}_{t^-}\right)$$

Note that we also have to include a starting time, s, as not all patients will start in the same state as with CRs. This also means we can define transition specific hazard functions as

$$\lambda_{hj}(t) = \lim_{\Delta t \to 0} \frac{\mathcal{P}_{hj}(t, t + \Delta t)}{\Delta t}$$

which has the same meaning as with CRs. It is the hazard of moving to state j which is applied to a patient whilst in state h. With this definition, we can give a mathematical definition of an absorbing state. If h is an absorbing state then $h \in \mathcal{S}$ such that $\forall t \in \mathcal{T}, j \in \mathcal{S}$, we have: $\lambda_{hj}(t) = 0$. Even in transient states, some transition intensities will be 0 for all t, implying that a transition can never occur (such as death to alive). When using multi-state modeling techniques it is clear that different covariates can induce a different effect on each transition, this includes cluster effects[65]. If all transitions in a model are one-way, then it is called a unidirectional, if one or more transitions can be reversed, then it is called reversible[105].

Let $\pi_h(0) = \text{Prob}\,(T(0) = h)$, i.e. the probability that the initial state (at $t = 0$) is state h. The probability of being in a given state at time t is therefore $\pi_h(t) = \sum_{j \in \mathcal{S}} \pi_j(0)\mathcal{P}_{jh}(0,t)$, which is the probability of being in each of the initial states, multiplied the probability of moving from that state to the current state (including the probability of starting in state h and staying there, $\pi_h(0)\mathcal{P}_{hh}(0,t)$).

The hazards of the transitions in an MSM can be represented by a matrix, where the row represents the previous state and the column is the new state, diagonal cells represent the "hazard" of staying in a given state and 0s imply that the transition does not occur. Note that the rows all add to 0 as the "hazard" being applied to a patient in a given state must be in equilibrium whilst the patient is there. Example from Anwar and Mahmoud[17]:

$$\begin{bmatrix} -\lambda_{12} - \lambda_{15} & \lambda_{12} & 0 & 0 & \lambda_{15} \\ 0 & -\lambda_{23} - \lambda_{25} & \lambda_{23} & 0 & \lambda_{25} \\ 0 & 0 & -\lambda_{34} - \lambda_{35} & \lambda_{34} & \lambda_{35} \\ 0 & 0 & 0 & -\lambda_{45} & \lambda_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

With transition hazards being expressed as a matrix, the probabilities of moving from one state to another between two given time points can similarly be expressed as a matrix

$$\mathcal{P}(s,t) = \begin{bmatrix} \mathcal{P}_{11}(s,t) & \mathcal{P}_{12}(s,t) & \mathcal{P}_{13}(s,t) & \mathcal{P}_{14}(s,t) & \mathcal{P}_{15}(s,t) \\ \mathcal{P}_{21}(s,t) & \mathcal{P}_{22}(s,t) & \mathcal{P}_{23}(s,t) & \mathcal{P}_{24}(s,t) & \mathcal{P}_{25}(s,t) \\ \mathcal{P}_{31}(s,t) & \mathcal{P}_{32}(s,t) & \mathcal{P}_{33}(s,t) & \mathcal{P}_{34}(s,t) & \mathcal{P}_{35}(s,t) \\ \mathcal{P}_{41}(s,t) & \mathcal{P}_{42}(s,t) & \mathcal{P}_{43}(s,t) & \mathcal{P}_{44}(s,t) & \mathcal{P}_{45}(s,t) \\ \mathcal{P}_{51}(s,t) & \mathcal{P}_{52}(s,t) & \mathcal{P}_{53}(s,t) & \mathcal{P}_{54}(s,t) & \mathcal{P}_{55}(s,t) \end{bmatrix}$$

At any time, t, we will have an event history, $\mathcal{H}_t$, which comes from a σ-algebra generated our function $T : \mathcal{T} \to \mathcal{S}$. The event history of a patient at time t is an element of this σ-algebra which consists of the observations of the process on the interval $[0, t]$. Multi-state models can be split into two approaches. The nonhomogeneous Markov approach assumes that $t = 0$ is when the patient enters the study and that time is tracked continuously across transitions. The Semi-Markov approach resets the time to $t = 0$ each time a new state is entered[65]. Ideally, clinical knowledge should be used to decide between these two approaches, however there are statistical methods that can assist in this decision[106,113]. It can useful to apply a weight to the sojourn times in a patient's history depending on the cost to the healthcare provider of being in that state[17]. Researchers should also be wary that in some cases, misclassification of a patient's state can occur, these are called Hidden Markov Models[106,114].

The Markov property means that the transition out of a state depends only on the current state and not any previously visited states[106], that is: $\mathcal{H}_t$ is ignored and the only part of the patient's history that is used is the current state of the patient and how long they have been there[105]. The Markov and Semi-Markov models can be refered to as "clock reset" and "clock forward" respectively due to the idea of resetting the clock when a patient enters a new state. When working under the "clock forward" ideology, MSMs will involve the evaluation of left-truncated data[106].

Although SDI could be considered to be a continuous or categorical variable, Bruce at al[50] used the fact that SDI is permanently increasing to model an MSM with the integer-valued SDI as an indicator of a patient's state (grouping all patients with SDI ≥ 5 into a single state for simplicity), see Figure 4, and see how that affects their mortality as well as a quality of life measure. Since SDI is a representation of permanent damage, this model is unidirectional MSM[50]. Their model used the Markov assumption that patient's state history did not affect the patient's current status and assumed a constant baseline hazard function for each of their transitions, $\lambda_{ij}(t) = \lambda_{0ij} \exp\left(\beta_{ij}^T T(t)\right)$. The results showed that, as would be expected, probability of death increased with higher SDI[50]. However it is important to note that it is not always possible to know the exact time of a transition leading to interval censoring (e.g. patient is in state A at $t = 10$ but is in state B at $t = 20$)[115]. This model developed by Bruce at al[50] assumed that patients can transition from the state SDI $= n$ to the state SDI $= n + 1$, but there is no mention of patients being able to skip over states (e.g. straight from SDI $= 1$ to SDI $= 3$) which could potentially happen often where a patient is only assessed once a year. Without these extra transitions, there would need to be an assumption that patients who appear to skip out steps actually move through them between check ups which implies some sort of interval censoring[115].

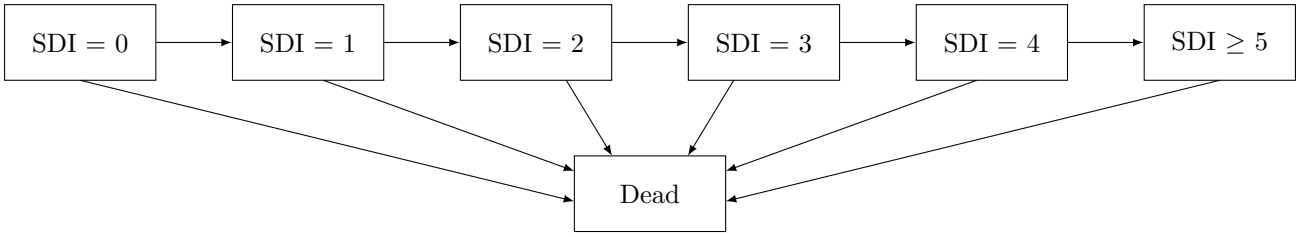

Figure 4: The progressive SDI model used by Bruce at al, 2015

A common MSM is the illness-death model, which has three states: healthy, illness and death[105]. Sometimes these states have different names and the "illness" state does not always actually imply a worse prognosis. The basic idea of this model is shown in Figure 3. In the illness-death model, we have three or four hazard functions depending on whether the transition from healthy to illness is reversible. Having only a few hazard functions makes this a relatively simple model, which is why it is quite common[105]. In a non-reversible illness-death model, we can define the transition probabilities depending on the three transitions intensities[105,106]:

$$\begin{aligned}
\mathcal{P}_{00}(s,t) &= \exp\left(-\int_s^t \lambda_{01}(u) + \lambda_{02}(u)\,\mathrm{d}u\right)\\
\mathcal{P}_{01}(s,t) &= \int_s^t \mathcal{P}_{00}(s,u-)\lambda_{01}(u)\mathcal{P}_{11}(u,t)\,\mathrm{d}u\\
\mathcal{P}_{11}(s,t) &= \exp\left(-\int_s^t \lambda_{12}(u)\,\mathrm{d}u\right)\\
\mathcal{P}_{12}(s,t) &= \int_s^t \mathcal{P}_{11}(s,u-)\lambda_{12}(u)\,\mathrm{d}u\\
\mathcal{P}^1_{02}(s,t) &= \int_s^t \mathcal{P}_{00}(s,u-)\lambda_{02}(u)\,\mathrm{d}u\\
\mathcal{P}^2_{02}(s,t) &= \int_s^t \mathcal{P}_{01}(s,u-)\mathcal{P}_{12}(u,t)\,\mathrm{d}u\\
\mathcal{P}_{02}(s,t) &= \mathcal{P}^1_{02}(s,t) + \mathcal{P}^2_{02}(s,t)
\end{aligned}$$

In these definitions, $\mathcal{P}^1_{02}$ is the probability of moving from state 0 to state 2 directly, whereas $\mathcal{P}^2_{02}$ is the probability of moving from state 0 to state 2 via state 1. In order to calculate the total probability of moving from state 0 to state 2, we have to sum these two probabilities. Similar to an earlier example, in $\mathcal{P}_{01}$, we have the probability of staying in state 0 from s to $u-$ multiplied by the probability of transitioning at time u and then by the probability of staying in state 1 from u to t integrated over all of u between s and t, this is quite an intuitive definition.

Another common model involves having two conditions, A & B, which may or may not be independent[105]. A simple way to model these two stages is as four states: No conditions, A only, B only or A & B. If the order of developing A and B is important, then the state A & B can be split into two states: A then B and B then A. See Figure 5.

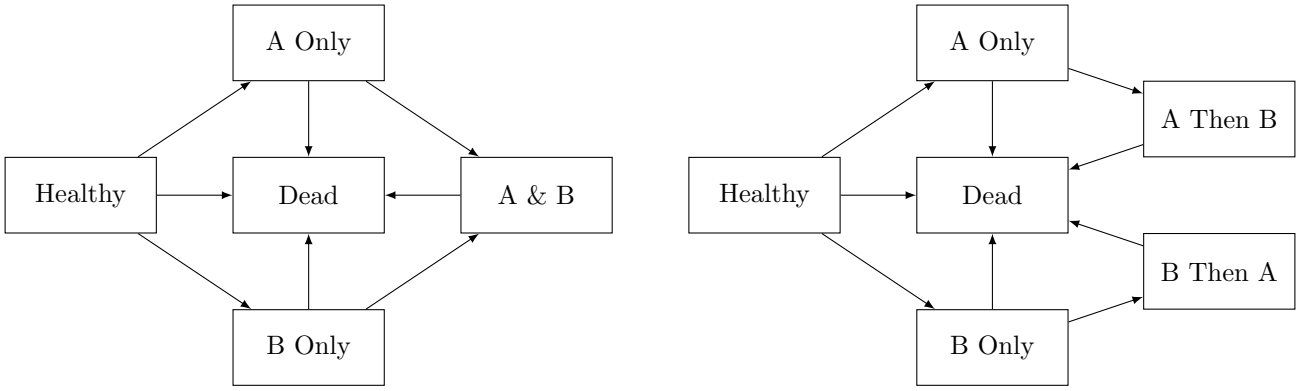

Figure 5: The A & B Models

MSMs can be used in conjunction with frailty modeling to produce robust models which can also handle

left truncation, right censoringand clustering[65]. Liquet et al[65] developed a multistate frailty model of patient's developing a Ventilator-associated pneumonia infection (VAP), being discharged or dying. They found that there was significant heterogeneity in the development of VAP and in discharge, but not significant heterogeneity in dying with or without VAP and discharge with VAP. This demonstrates that clustering can occur in some transitions but not in others and provides evidence that frailty should always be included in a large scale study. This study concluded that VAP is frequent in Intensive Care Units (ICUs) and is associated with an increase in ICU mortality, length of stay and cost.

Anwar and Mahmoud[17] developed a stochastic model which treated CKD progression as a series of states in an MSM with three state being dependent on the Glomerular Filtration Rate (GFR) of the kidneys (mild, moderate and severe reduction) and ESRD being the fourth stage and death being the final, absorbing state. They used this model to estimate survival probability, which does not require detailed knowledge of the transient states. This turns the all "non-dead" states into sub-states of a larger "living" state, see Figure 6. Although prediction for patients should still always use the relevant hazard function for their current state, this kind of idea allows average times to be infered from the data, as is done in their paper.

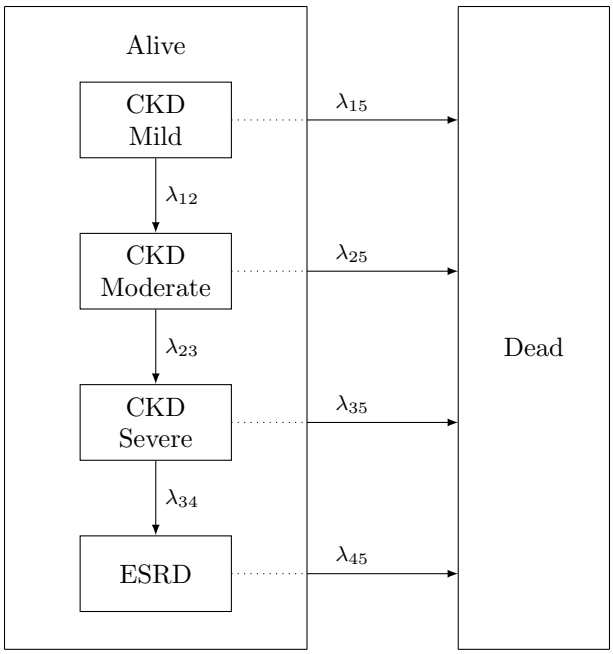

Figure 6: Transition from Alive to Dead from Anwar and Mahmoud, 2014[17]

CRs and MSMs are versatile mechanics which can be used to great advantage of rearchers if done properly. By mapping the journey through the system of every patient in a population, detailed predictions can be made for future patients. As well as being able to provide a patient with the probability of them having a particular event, we can also tell the patient the probability of them being in a given state in 1/5/10 years time. The probability of transitioning through a state can also be provided. MSMs can become very complicated very quickly, and even something as simple as adding reversibility to a transition increases the convolutions that can occur. As discussed in Section 2, MSMs lead themselves very well to the transitions taking during RRT. Clearly this model may be complicated, but as long as the transparency of development and validation is preserved, the end result could be a very useful and usable clinical prediction model.

5 Conclusion

There is currently a wealth of evidence that there are methodological shortcomings in the current prognostic research field[1–4,12,34–36,40,77,116–120] to a point where there is almost a saturation of cynicism there. It is clear that there needs to be more guidance in the field[1], but this guidance needs to be channeled into education rather than criticism. An improvement in education of prognostic research could have a dramatic effect on the usability of research data for prognostic research. Teaching people how to generate and use this kind of data as undergraduates, postgraduates and healthcare workers would greatly improve data quality and allow researchers to focus on good methodology rather than fixing data quality issues. One common occurance which could be improved is that most prognostic studies are retrospective[40], whereas it would be preferable to use data from a prospective study.

It is also common that the concerns of patients are not always in line with the research being performed, for example, patients are more interested in their symptoms than the causes of those symptoms and although reducing the cause often reduces the symptoms, this is not always the case[121,122]. With improvements to the methodology being used, we could also see models being used more commonly in clinical practice as it is a research field which is currently massively under utilised in the real-world[1]. This is due, in part, to the fact that at the current time, prognostic research is subsumed and fragmented into other fields of research[1]. It is clear that more cohesion amongst prognostic researchers would improve the entire field and have wider implications. A simple improvement in the rigours of scientific methods used in research could have a drastic improvement to the quality of predictive models produced and a reduction in the quantity of poor ones. However, as researchers priorities are often skewed towards "novel and interesting" results rather than thorough and needed ones, this might be a long time coming[95].

With big data becoming more and more common, validation studies should become easier to come by. Hopefully this access to big data will be augmented by the (relatively young) TRIPOD guidelines[37,38]. By sticking to the guidelines when reporting their data, model developers can allow potential model validators to understand and repeat their methods enough to provide an adequate external validation (which should then also follow the TRIPOD guidelines). Depending on the area of study, other guidelines similar to TRIPOD may be useful to researchers such as the REMARK checklist for prognostic factor research[123] and the GRIPS statement for genomic risk factors[124] which were actually integral in it's formation[37,38].

Although MSMs are an uncommon method of modeling patient pathways, it is clear that they are useful. As discussed in Section 2, CKD is an important health factor for future populations and it is clearly amenable to the idea of MSMs. Section 3 instills the idea that when producing a prognostic model, there are many aspects to consider with transparency of research being a clear contender for most important. Section 4 discussed the theory behind what an MSM is and how it can be utilised to produce dynamic models for any disease which can be divided into states. Overall, this report indicates that although there is a wealth of knowledge in the literature regarding these three topics and even though they have been combined together in some instances, their combination in this manner is novel and will hopefully produce a viable and transparent multi-state model which can easily be validated, assessed for impact and implemented by nephrologists as a decision making tool.

References

[1] Harry Hemingway et al. "Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes." In: *BMJ (Clinical research ed.)* 346.February 2013 (2013), e5595. ISSN: 1756-1833. DOI: 10.1136/bmj.e5595. URL: http://www.ncbi.nlm.nih.gov/pubmed/23386360%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3565687.

[2] Richard D Riley et al. "Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research". In: *PLoS Medicine* 10.2 (Feb. 2013), e1001380. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001380. URL: http://www.ncbi.nlm.nih.gov/pubmed/23386360%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3565687%20http://dx.plos.org/10.1371/journal.pmed.1001380.

[3] Ewout W. Steyerberg et al. "Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research". In: *PLoS Medicine* 10.2 (Feb. 2013), e1001381. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001381. URL: http://dx.plos.org/10.1371/journal.pmed.1001381.

[4] Aroon D Ad Hingorani et al. "Prognosis research strategy (PROGRESS) 4: Stratified medicine research". In: *Bmj* 5793.February 2013 (2013), pp. 1–9. ISSN: 1756-1833. DOI: 10.1136/bmj.e5793. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3565686%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%7B%5C%%7D5Cnhttp://www.bmj.com/content/346/bmj.e5793.full.

[5] J L Haybittle et al. "A prognostic index in primary breast cancer." In: *British journal of cancer* 45.3 (1982), pp. 361–6. ISSN: 0007-0920. DOI: 10.1038/bjc.1982.62. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2010939%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[6] Julia Hippisley-Cox et al. "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study". In: *BMJ* 335 (2007). DOI: 10.1136/bmj.39261.471806.55.

[7] Julia Hippisley-Cox et al. "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2." In: *Bmj* 336.7659 (2008), pp. 1475–1482. ISSN: 0959-8138. DOI: 10.1136/bmj.39609.449676.25.

[8] Julia Hippisley-Cox et al. "Advantages of QRISK2 (2010): the key issue is ethnicity and extent of reallocation". In: *Heart* 97.6 (2011), p. 515. DOI: 10.1136/hrt.2010.221085.

[9] QRESEARCH®. *QRISK®2*. https://qrisk.org. 2016.

[10] Gary S Collins and Douglas G Altman. "Predicting the adverse risk of statin treatment: an independent and external validation of Qstatin risk scores in the UK." In: *Heart (British Cardiac Society)* 98.14 (2012), pp. 1091–7. ISSN: 1468-201X. DOI: 10.1136/heartjnl-2012-302014. URL: http://www.ncbi.nlm.nih.gov/pubmed/22689714.

[11] Robert C Atkins. "The epidemiology of chronic kidney disease". In: *Kidney International* 67 (2005), S14–S18.

[12] Gary S. Collins et al. "A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods". In: *Journal of Clinical Epidemiology* 66.3 (2013), pp. 268–277. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2012.06.020. URL: http://dx.doi.org/10.1016/j.jclinepi.2012.06.020.

[13] Katherine T Mills et al. "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010". In: *Kidney International* 88 (2015), pp. 950–957.

[14] Julie Gilg, Fergus Casker, and Damian Fogarty. "UK Renal Registry 18th Annual Report: Chapter 1 UK Renal Replacement Therapy Incidence in 2014: National and Centre-specific Analysis". In: *Nephron* 132.suppl.1 (2016), pp. 9–40. DOI: 10.1159/000444815.

[15] David Ansell et al. "The Ninth Annual Report". In: *UK Renal Registry Report* (2006).

[16] National Institutes of Health et al. "US Renal Data System, USRDS 2010 Annual data report: atlas of chronic kidney disease and end-stage renal disease in the United States". In: *National Institute of Diabetes and Digestive and Kidney Diseases* (2010).

[17] Noura Anwar and Mahmoud Riad Mahmoud. "A Stochastic Model for the Progression of Chronic Kidney Disease". In: *ijera* 4.11 (2014), pp. 8–19.

[18] Manjula Kurella Tamura, Jane C. Tan, and Ann M. O'Hare. "Optimizing renal replacement therapy in older adults: a framework for making individualized decisions". In: *Kidney International* 82.3 (2012), pp. 261–269. ISSN: 00852538. DOI: 10.1038/ki.2011.384. URL: http://linkinghub.elsevier.com/retrieve/pii/S0085253815555324.

[19] Ann M O'Hare et al. "Regional Variation in Health Care Intensity and Treatment Practices for End-Stage Renal Disease in Older Adults". In: *JAMA* 304.2 (2010), pp. 180–186.

[20] Ya-Chen Tina Shih et al. "Impact of initial dialysis modality and modality switches on Medicare expenditures of end-stage renal disease patients". In: *Kidney international* 68 (2005), pp. 319–329.

[21] Robert N Foley, Shu-Cheng CHEN, and Allan J Collins. "Hemodialysis access at initiation in the United States, 2005 to 2007: still "catheter first"". In: *Hemodialysis International* 13.4 (2009), pp. 533–542. DOI: 10.1111/j.1542-4758.2009.00396.x.

[22] Vascular Access Work Group. "Clinical practice guidelines for vascular access". In: *American Journal of Kidney Diseases* 48.1 Suppl 1 (2006), S248–S257. DOI: 10.1053/j.ajkd.2006.04.040.

[23] AI Richardson 2nd et al. "Should fistulas really be first in the elderly patient?" In: *The Journal of Vascular Access* 10.3 (2009), pp. 199–202.

[24] Jeffrey Perl et al. "Hemodialysis Vascular Access Modifies the Association between Dialysis Modality and Survival". In: *Journal of the American Society of Nephrology* 22 (2011), pp. 1113–1121. DOI: 10.1681/ASN.2010111155.

[25] Rajnish Mehrotra et al. "Similar Outcomes with Hemodialysis and Peritoneal Dialysis in Patients With End-Stage Renal Disease". In: *Archives of Internal Medicine* 171.2 (2011), pp. 110–118. DOI: 10.1001/archinternmed.2010.352.

[26] UCLA Health. *UCLA Kidney Exchange*. Online at: transplants.ucla.edu/kidneyexchange.

[27] Matthew Cooper and Cynthia L Forland. "The elderly as recipients of living donor kidneys, how old is too old?" In: *Current Opinion in Organ Transplantation* 16.2 (2011), pp. 250–255.

[28] Robert A Wolfe et al. "Comparison of Mortality in All Patients on Dialysis, Patients on Dialysis Awaiting Transplantation, and Recipients of a First Cadaveric Transplant". In: *New England Journal of Medicine* 341.23 (1999), pp. 1725–1730.

[29] Robert M Merion et al. "Deceased-Donor Characteristics and the Survival Benefit of Kidney Transplantation". In: *Jama* 294.21 (2005), pp. 2726–2733.

[30] Sarbjit V Jassal et al. "Kidney transplantation in the elderly: a decision analysis". In: *Journal of the American Society of Nephrology* 14 (2003), pp. 187–196. DOI: 10.1097/01.ASN.0000042166.70351.57.

[31] Hippocrates and Francis Adams. "On Airs, Waters and Places". In: *The Genuine Works of Hippocrates*. Sydenham Society, London, 1849, pp. 179–222.

[32] Kristian Thygesen, Joseph S Alpert, and Harvey D White. "Universal Definition of Myocardial Infarction". In: *Journal of the American College of Cardiology* 50.22 (2007), pp. 2173–2195. DOI: 10.1016/j.jacc.2007.09.011.

[33] Vincent Probst et al. "Long-Term Prognosis of Patients Diagnosed with Brugada Syndrome Results From the FINGER Brugada Syndrome Registry". In: *Circulation* 121 (2010), pp. 635–643. DOI: 10.1161/CIRCULATIONAHA.109.887026.

[34] Patrick Royston et al. "Prognosis and prognostic research: Developing a prognostic model". In: *BMJ (Online)* 338.mar31 1 (Mar. 2009), b604–b604. ISSN: 0959-8138. DOI: 10.1136/bmj.b604. URL: http://www.bmj.com/cgi/doi/10.1136/bmj.b604.

[35] Douglas G Altman et al. "Prognosis and prognostic research: validating a prognostic model." In: *BMJ (Clinical research ed.)* 338.june (2009), b605. ISSN: 0959-8138. DOI: 10.1136/bmj.b605.

[36] Karel G M Moons et al. "Prognosis and prognostic research: application and impact of prognostic models in clinical practice." In: *BMJ (Online)* 338.6 (2009), b606. ISSN: 1756-1833. DOI: 10.1136/bmj.b606. URL: http://www.ncbi.nlm.nih.gov/pubmed/19502216.

[37] Gary S. Collins et al. "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement". In: *Annals of Internal Medicine* 162.1 (Jan. 2015), p. 55. ISSN: 0003-4819. DOI: 10.7326/M14-0697. URL: http://annals.org/article.aspx?doi=10.7326/M14-0697.

[38] Karel G.M. Moons et al. "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration". In: *Annals of Internal Medicine* 162.1 (2015), W1. ISSN: 0003-4819. DOI: 10.7326/M14-0698.

[39] Richard D Riley et al. "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges". In: *Bmj* 353.mar31_1 (2016), p. i3140. ISSN: 1756-1833. DOI: 10.1136/bmj.i3140. URL: http://www.ncbi.nlm.nih.gov/pubmed/27334381%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4916924%7B%5C%%7D5Cnhttp://www.bmj.com/lookup/doi/10.1136/bmj.i3140.

[40] Karel G M Moons et al. "Prognosis and prognostic research : what, why, and how ?" In: *BMJ (Online)* 375.February (2009), pp. 1–7. DOI: 10.1136/bmj.b375.

[41] The Academy of Medical Sciences. *Realising the potential of stratified medicine*. July 2013.

[42] Mark R Trusheim, Ernst R Berndt, and Frank L Douglas. "Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers". In: *Nature Reviews Drug Discovery* 6 (2007), pp. 287–293. DOI: 10.1038/nrd2251.

[43] Tyler Vigen. *Spurious Correlations*. Hachette Books, 2015. ISBN: 0316339431.

[44] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea". In: *Statistics in Medicine* 25 (2006), pp. 127–141. DOI: 10.1002/sim.2331.

[45] The International Warfarin Pharmacogenetics Consortium. "Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data". In: *New England Journal of Medicine* 360.8 (2009), pp. 753–764. DOI: 10.1056/NEJMoa0809329.

[46] Andrew J Vickers, Ethan Basch, and Michael W Kattan. "Against diagnosis". In: *Ann Intern Med* 149 (2008), pp. 200–203. DOI: 149/3/200[pii].

[47] Danilo Fliser et al. "Fibroblast Growth Factor 23 (FGF23) Predicts Progression of Chronic Kidney Disease: The Mild to Moderate Kidney Disease (MMKD) Study". In: *J Am Soc Nephrol* 18 (2007), pp. 2600–2608. DOI: 10.1681/ANS.2006080936.

[48] Maria GM Hunink et al. "The recent decline in mortality from coronary heart disease, 1980-1990: the effect of secular trends in risk factors and treatment". In: *JAMA* 277.7 (1997), pp. 535–542.

[49] Dafna D Gladman et al. "The Reliability of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index in Patients with Systemic Lupus Erythematosus". In: *Arthritis & Rheumatism* 40.5 (1997), pp. 809–813.

[50] Ian N Bruce et al. "Factors associated with damage accrual in patients with systemic lupus erythematosus: results from the Systemic Lupus International Collaborating Clinics (SLICC) Inception Cohort." In: *Annals of the rheumatic diseases* 74 (2015), pp. 1706–1713. ISSN: 1468-2060. DOI: 10.1136/annrheumdis-2013-205171. URL: http://www.ncbi.nlm.nih.gov/pubmed/24834926.

[51] Nish Chaturvedi. "Ethnic Differences in Cardiovascular Disease". In: *Heart* 89 (2003), pp. 681–686.

[52] Paramjit S Gill et al. "Black and Minority Ethnic Groups". In: *Health Care Needs Assessment: The Epidemiogically Based needs Assessment reviews.* Ed. by Raferty J. Abingdon. Radcliffe Medical Press Ltd, 2007. Chap. 4, pp. 227–399.

[53] Richard D Riley et al. “A systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing’s sarcoma and neuroblastoma”. In: *Health Technology Assessment* 7.5 (2003).

[54] S. E. Bleeker et al. “External validation is necessary in prediction research: A clinical example”. In: *Journal of Clinical Epidemiology* 56.9 (2003), pp. 826–832. ISSN: 08954356. DOI: 10.1016/S0895-4356(03)00207-5.

[55] Justin Zaman and Eric Brunner. “Social inequalities and cardiovascular disease in South Asians”. In: *Heart* 94.4 (2008), pp. 406–407.

[56] David A Hanauer, Daniel R Rhodes, and Arul M Chinnaiyan. “Exploring Clinical Associations Using ’-Omics’ Based Enrichment Analyses”. In: *PloS ONE* 4.4 (2009), e5203. DOI: 10.1371/journal.pone.0005203.

[57] Peter Peduzzi et al. “Importance of Events per Independent Variable in Proportional Hazards Regression Analysis II. Accuracy and Precision of Regression Estimates”. In: *Journal of Clinical Epidemiology* 48.12 (1995), pp. 1503–1510.

[58] Peter Peduzzi et al. “A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis”. In: *Journal of Clinical Epidemiology* 49.12 (1996), pp. 1373–1379.

[59] Willi Sauerbrei, Patrick Royston, and Harald Binder. “Selection of important variables and determination of functional form for continuous predictors in multivariable model building”. In: *Statistics in Medicine* 26.30 (Dec. 2007), pp. 5512–5528. ISSN: 02776715. DOI: 10.1002/sim.3148. arXiv: NIHMS150003. URL: http://doi.wiley.com/10.1002/sim.3148.

[60] Yvonne Vergouwe et al. “Substantial effective sample sizes were required for external validation studies of predictive logistic regression models”. In: *Journal of Clinical Epidemiology* 58 (2005), pp. 475–483. DOI: 10.1016/j.jclinepi.2004.06.017.

[61] Gary S Collins, Emmanuel O Ogundimu, and Douglas G Altman. “Sample size considerations for the external validation of a multivariable prognostic model: a resampling study”. In: *Statistics in Medicine* 35 (2016), pp. 214–226. DOI: 10.1002/sim.6787.

[62] C Counsell and M Dennis. “Systematic review of prognostic models in patients with acute stroke”. In: *Cerebrovascular Diseases* 12.3 (2001), pp. 159–170. DOI: http://dx.doi.org/10.1159/000047699. URL: http://easyaccess.lib.cuhk.edu.hk/login?url=http://ovidsp.ovid.com/ovidweb.cgi?T=JS%7B%5C&%7DCSC=Y%7B%5C&%7DNEWS=N%7B%5C&%7DPAGE=fulltext%7B%5C&%7DD=med4%7B%5C&%7DAN=11641579%7B%5C%7D5Cnhttp://findit.lib.cuhk.edu.hk/852cuhk/?sid=OVID:medline%7B%5C&%7Did=pmid:11641579%7B%5C&%7Did=doi:%7B%5C&%7Dissn=1015-9770%7B%5C&%7Disbn=%7B%5C&%7Dvolume=12%7B%5C&%7Dissue=3%7B%5C&%7Dspag.

[63] Joan Ivanov et al. “Predictive Accuracy Study: Comparing a Statistical Model to Clinicians’ Estimates of Outcomes After Coronary Bypass Surgery”. In: *The Annals of thoracic surgery* 70 (2000), pp. 162–168.

[64] J H Todd et al. “Confirmation of a prognostic index in primary breast cancer.” In: *British journal of cancer* 56.4 (1987), pp. 489–92. ISSN: 0007-0920. DOI: 10.1038/bjc.1987.230. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2001834%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[65] Benoit Liquet, Jean Francois Timsit, and Virginie Rondeau. “Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units”. In: *BMC Medical Research Methodology* 12.1 (2012), p. 79. ISSN: 1471-2288. DOI: 10.1186/1471-2288-12-79.

[66] Kym IE Snell et al. “Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model”. In: *Journal of Clinical Epidemiology* 69 (2016), pp. 40–50. DOI: 10.1016/j.jclinepi.2015.05.009.

[67] Philip Hougaard. “Frailty Models for Survival Data”. In: *Lifetime Data Analysis* 1.3 (), pp. 255–273. DOI: 10.1007/BF00985760.

[68] A Rogier T Donders et al. “Review: A gentle introduction to imputation of missing values”. In: *Journal of Clinical Epidemiology* 59.10 (2006), pp. 1087–1091. DOI: 10.1016/j.jclinepi.2006.01.014.

[69] Frank Harrell. *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer-Verlag New York, 2001. ISBN: 978-1-4419-2918-1. DOI: 10.1007/978-1-4757-3462-1.

[70] S W Lagakos. “Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable”. In: *Statistics in Medicine* 7.1-2 (1988), pp. 257–274.

[71] L H J Eberhart et al. “Applicability of risk scores for postoperative nausea and vomiting in adults to paediatric patients”. In: *British Journal of Anaesthesia* 93.3 (2004), pp. 386–392.

[72] RD Riley et al. “Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future”. In: *British Journal of Cancer* 88 (2003), pp. 1191–1198. DOI: 10.1038/sj.bjc.6600886.

[73] Cox R David. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society* 34.2 (1972), pp. 187–220.

[74] Nathan Mantel. “Why stepdown procedures in variable selection”. In: *Technometrics* 12.3 (1970), pp. 621–625.

[75] Willi Sauerbrei. “The use of resampling methods to simplify regression models in medical statistics”. In: *Applied Statistics* 48.3 (1999), pp. 313–329.

[76] Patrick Royston, Gareth Ambler, and Willi Sauerbrei. “The use of fractional polynomials to model continuous risk variables in epidemiology.” In: *International journal of epidemiology* 28 (1999), pp. 964–974.

[77] Susan Mallett et al. “Reporting methods in studies developing prognostic models in cancer: a review”. In: *BMC medicine* 8 (2010), p. 20.

[78] Walter Bouwmeester et al. “Reporting and methods in clinical prediction research: a systematic review”. In: *PLoS Med* 9.5 (2012), e1001221. DOI: 0.1371/journal.pmed.1001221.

[79] Alvin H Moss et al. “Utility of the "Surprise" Question to Identify Dialysis Patients with High Mortality”. In: *Clinical Journal of the American Society of Nephrology* 3 (2008), pp. 1379–1384. DOI: 10.2215/CJN.00940208.

[80] Brendan M Reilly and Arthur T Evans. “Translating clinical research into clinical practice: impact of using prediction rules to make decisions”. In: *Annals of Internal Medicine* 144 (2006), pp. 201–209.

[81] QRESEARCH. *What Is QRESEARCH®?* http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx. 2012.

[82] Frank E Harrell, Kerry L Lee, and Daniel B Mark. “Tutorial in Biostatistics: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors”. In: *Statistics in medicine* 15 (1996), pp. 361–387.

[83] William J Mackillop and Carol F Quirt. “Measuring the Accuracy of Prognostic Judgments in Oncology”. In: *Journal of clinical epidemiology* 50 (1997), pp. 21–29.

[84] Patrick Royston and Willi Sauerbrei. “A new measure of prognostic separation in survival data”. In: *Statistics in Medicine* 23 (2004), pp. 723–748. DOI: 10.1002/sim.1621.

[85] Michael J Pencina, Ralph B D’Agostino, and Ramachandran S Vasan. “Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond”. In: *Statistics in Medicine* 27 (2008), pp. 157–172. DOI: 10.1002/sim.2929.

[86] Youlian Liao et al. “How generalizable are coronary risk prediction models? Comparison of Framingham and two national cohorts”. In: *American heart journal* 137 (1999), pp. 837–845.

[87] Paolo Fraccaro et al. “An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK”. In: *BMC Medicine* 14.1 (2016), p. 104. ISSN: 1741-7015. DOI: 10.1186/s12916-016-0650-2. URL: http://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0650-2.

[88] David W Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2000. ISBN: 9780471356325. DOI: 10.1002/0471722146.

[89] Ewout W Steyerberg et al. “Internal validation of predictive models: efficiency of some procedures for logistic regression analysis”. In: *Journal of Clinical Epidemiology* 54 (2001), pp. 774–781.

[90] Amy C Justice, Kenneth E Covinsky, and Jesse A Berlin. “Assessing the Generalizability of Prognostic Information”. In: *Annals of Internal Medicine* 130 (1999), pp. 515–524.

[91] J André Knottnerus. “Between iatrotropic stimulus and interiatric referral:: the domain of primary care research”. In: *Journal of Clinical Epidemiology* 55 (2002), pp. 1201–1206.

[92] John H Wasson et al. “Clinical prediction rules: applications and methodological standards”. In: *New England Journal of Medicine* 313.13 (1985), pp. 793–799.

[93] Andreas Laupacis, Nandita Sekar, et al. “Clinical prediction rules: a review and suggested modifications of methodological standards”. In: *JAMA* 277.6 (1997), pp. 488–494.

[94] Sonja Grill et al. “Comparison of approaches for incorporating new information into existing risk prediction models”. In: *Statistics in Medicine* (2016). DOI: 10.1002/sim.7190.

[95] Derek Muller for Veritasium. *Is Most Published Research Wrong?* Online Video. YouTube. Published 11 Aug 2016. Accessed 19 Dec 2016. https://www.youtube.com/watch?v=42QuXLucH3Q.

[96] Angela Cooper and Norma O'Flynn. "Guidelines: Risk Assessment and Lipid Modification for Primary and Secondary Prevention of Cardiovascular Disease: Summary of NICE Guidance". In: *BMJ* 336.7655 (2008), pp. 1246–1248.

[97] NICE. *Cardiovascular disease: risk assessment and reduction, including lipid modification.* https://www.nice.org.uk/guic recommendations. Accessed:19th December 2016. July 2014.

[98] Keaven M Anderson et al. "Cardiovascular disease risk profiles". In: *American Heart Journal* 121.1 (1991), pp. 293–298.

[99] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987. ISBN: 0-471-08705-X.

[100] Julia Hippisley-Cox and Carol Coupland. "Predicting the risk of Chronic Kidney Disease in Men and Women in England and Wales: prospective derivation and external validation of the QKidney® Scores". In: *BMC Family Practice* 11.1 (2010), p. 49.

[101] Ian Stiell et al. "Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries". In: *BMJ* 311.7005 (1995), p. 594. DOI: 10.1136/bmj.311.7005.594.

[102] Cathy Cameron and C David Naylor. "No impact from active dissemination of the Ottawa Ankle Rules: further evidence of the need for local implementation of practice guidelines". In: *Canadian Medical Association Journal* 160.8 (1999), pp. 1165–1168.

[103] Marion K Campbell, Diana R Elbourne, and Douglas G Altman. "CONSORT statement: extension to cluster randomised trials". In: *BMJ* 328 (2004), pp. 702–708.

[104] Jonathan C Hill et al. "Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial". In: *The Lancet* 378.9802 (2011), pp. 1560–1571. DOI: 10.1016/S0140-6736(11)60937-9.

[105] Per Kragh Andersen and Niels Keiding. "Multi-state models for event history analysis." In: *Statistical methods in medical research* 11.2 (2002), pp. 91–115. ISSN: 09622802. DOI: 10.1191/0962280202SM276ra.

[106] H. Putter, M. Fiocco, and R. B. Geskus. "Tutorial in biostatistics: competing risks and multi-state models". In: *Statistics in Medicine* 26.11 (May 2007), pp. 2389–2430. ISSN: 02776715. DOI: 10.1002/sim.2712. arXiv: NIHMS150003. URL: http://doi.wiley.com/10.1002/sim.2712.

[107] David R Cox. "Partial likelihood". In: *Biometrika* 62.2 (1975), pp. 269–276.

[108] Per Kragh Andersen, Steen Z Abildstrom, and Susanne Rosthøj. "Competing risks as a multi-state model." In: *Statistical methods in medical research* 11.2 (2002), pp. 203–215. ISSN: 09622802. DOI: 10.1191/0962280202sm281ra.

[109] Daniel Bernoulli. "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir". In: *Histoire de l'Acad. Roy. Sci.(Paris) avec Mém. des Math. et Phys. and Mém* (1760), pp. 1–45.

[110] Daniel Bernoulli and Sally Blower. "An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it". In: *Reviews in medical virology* 14.5 (2004), pp. 275–288.

[111] Jason P Fine and Robert J Gray. "A Proportional Hazards Model for the Subdistribution of a Competing Risk". In: *Journal of the American Statistical Association* 94.446 (June 1999), pp. 496–509. ISSN: 01621459. DOI: 10.2307/2670170. URL: http://www.jstor.org/stable/2670170?origin=crossref.

[112] Sandrine Katsahian et al. "Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution". In: *Statistics in medicine* 25 (2006), pp. 4267–4278. DOI: 10.1002/sim.2684.

[113] Daniel Commenges et al. "Choice between Semi-parametric Estimators of Markov and Non-Markov Multi-state Models from Coarsened Observations". In: *Scandinavian Journal of Statistics* 34.1 (2007), pp. 33–52. DOI: 10.1111/j.1467-9469.2006.00536.x.

[114] Christopher H Jackson et al. "On Modelling Minimal Disease Activity". In: *Arthritis Care & Research* 68.3 (2016), pp. 388–393. DOI: 10.1002/acr.22687.

[115] Daniel Commenges. "Inference for multi-state models from interval-censored data". In: *Statistical Methods in Medical Research* 11.2 (2002), pp. 167–182.

[116] Catherine Meads, Ikhlaaq Ahmed, and Richard D Riley. "A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance". In: *Breast Cancer Research and Treatment* 132 (2012), pp. 365–377. DOI: 10.1007/s10549-011-1818-2.

[117] Douglas G Altman. "Prognostic models: A Methodological Framework and Review of Models for Breast Cancer". In: *Cancer investigation* 27.3 (2009), pp. 235–243. DOI: 10.1080/07357900802572110.

[118] Susan Mallett et al. "Reporting performance of prognostic models in cancer: a review". In: *BMC medicine* 8 (2010), p. 21.

[119] Gary S Collins et al. "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting". In: *BMC Medicine* 9.1 (2011), p. 103. DOI: 10.1186/1741-7015-9-103.

[120] Brian Buijsse et al. "Risk Assessment Tools for Identifying Individuals at Risk of Developing Type 2 Diabetes". In: *Epidemiologic Reviews* 33.1 (2011), pp. 46–64. DOI: 10.1093/epirev/mxq019.

[121] Harry Hemingway et al. "Evaluating the Quality of Research into a Single Prognostic Biomarker: A Systematic Review and Meta-analysis of 83 studies of C-Reactive Protein in Stable Coronary Artery Disease". In: *PLoS Med* 7.6 (2010), e1000286. DOI: 10.1371/journal.pmed.1000286.

[122] Han Repping-Wuts et al. "Fatigue as experienced by patients with rheumatoid arthritis (RA): a qualitative study". In: *International Journal of Nursing Studies* 45.7 (2008), pp. 995–1002. DOI: 10.1016/j.ijnurstu.2007.06.007.

[123] Lisa M McShane et al. "Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK)". In: *Journal of the National Cancer Institute* 97.16 (2005), pp. 1180–1184.

[124] A Cecile J W Janssens et al. "Strengthening the reporting of genetic risk prediction studies: the GRIPS statement". In: *Genome Medicine* 3.3 (2011), p. 16. DOI: 10.1186/gm230.