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

MA Barrowman GP Martin N Peek M Lambie M Sperrin

Contents

1 Introduction

Clinical Prediction Models (CPMs) provide individualised risk of a patient's outcome (cite), based on that patient's predictors. These predictions will usually be in the form of a risk score or probability. However, using traditional modelling techniques, these CPMs will only predict a single outcome. Multi-State Clinical

Prediction Models (MS-CPMs) combine the multi-state modelling framework to the prognostic field to provide predictions for multiple outcomes in a single model. Once a CPM has been developed, it is important to assess how well the model actually performs (cite). This process is called Model Validation and involves comparing the predictions produced by the model to the actual outcomes experienced by patients (cite). It is expected that the development of a CPM will be accompanied by the validation of the model on the same dataset it was developed in (internal validation), using either bootstrapping or cross-validation to account for optimism in the developed model (cite). Models can also be validated on a novel dataset (external validation), which is used to assess the generalisability and transportability of the model (cite). During validation, there are different aspects of model performance that we can assess and these are measured using specific metrics. For example, to assess the overall Accuracy of a model, we may use the Brier Score (cite) or to analyse how well a model discriminates between patients, we could use the c-statistic (cite). The current metrics that are commonly used have been designed and extended to work in a variety of model development frameworks. However, these extensions are limited to either a single outcome (as in traditionally developed models) or do not adequately account for the censoring of patients (as commonly occurs in longitudinal data). This paper aims to provide use-able extensions to current performance metrics to be used when validating MS-CPMs. It is essential that these extensions are directly comparable with current metrics (to allow for quicker adoption), that they are collapsible to the current metrics and that they adjust for the bias induced by the censoring of patients. Currently, the most common way to validate an MS-CPMs is by applying traditional methods to compare across two states at a given time and then aggregating the results in an arbitrary manner [cite something]. Other methodologists have extended existing metrics to multinomial outcomes [cite van Calster], which do not contain a time-based component; to simple competing risks scenarios [cite CR c-statistic], which do not contain transient states; or to [... insert third relevant example]. Spitoni et al [cite Spitoni 2018]] developed methods to apply the Brier Score (or any proper score functions) to a multi-state setting and so a simplified and specific version of their work is described in this paper. It is the hope of the authors that this work will increase the uptake of multi-state models and the sub-field of MS-CPMs will grow appropriately.

2 Motivating Data Set

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

Throughout this paper we will use a model developed in Chronic Kidney Disease (CKD) patients to assess their progression onto Renal Replacement Therapy (RRT) and/or Death [cite Dev/Valid Paper]. The model was developed using data from the Salford Kidney Study (SKS) and then applied to an external dataset derived from the West of Scotland (see Table 2) [1]. The original model predicts the probability that a patient has begun RRT and/or died after their first recorded eGFR below 60 ml/min/1.73m2, by any time in the future (reliable up to 10 years). For the purposes of this paper, we will take a snapshot of the predictions at the 5 year time point. The Three-State model used in our example is designed as an Illness-Death Model [2], this is one of the simplest MSM designs and has the key advantage over a traditional model that they can predict whether a patient is in or has visited the transient state before reaching the absorbing state (i.e. patient who became ill before dying or who started RRT before dying) (see figure 1).

[**Figure of the MSM**]

[**Describe Glasgow Data**]

3 Current Approaches

Here we describe three commonly used performance metrics for assessing the performance of a traditional survival clinical prediction model. These metrics assess the Accuracy, Discrimination and Calibration of the models being validated. Accuracy is an overall measurement of how well the model predicts the outcomes in the patients. Discrimination assesses how well the model discerns between patients; in a two-state model

this is a comparison of patients with and without the outcome, and should assign a higher value to those that experience the outcome. Calibration is the agreement between the observed outcomes and the predicted risks across the full risk-range. We are applying cross-sectional metrics at a set time point within the setting of a longitudinal model and so we need to account for the censoring of patients and therefore, each uncensored patient at a given time t will be weighted as per the Inverse Probability of Censoring Weighting (IPCW) [3]. This allows the uncensored patient population to be representative of the entire patient population.

3.1 Baseline Models

To assess the performance of a model, we must compare the values produced by the performance metrics to those of two baseline models; a random or noninformative model and a perfect model. A Non-Informative (NI-)model assigns the same probability to all patients to be in any state regardless of covariates and is akin to using the average prevalence in the entire population to define your model. For example, in a Two-State model and an event that occurs in 10% of patients, all patients are predicted to have a 10% chance of having the event. For many metrics, models can be compared to a Non-Informative model to assess whether the model is in fact "better than random". A Perfect (P-)model is one which successfully assigns a 100% probability to all patients, and the predictions are correct; this is the ideal case, which many models can also be compared to as models as close to this display excellent predictive abilities. Although models may perform worse than a non-informative one, we will not consider these in detail here as they are considered to be without worth in terms of predictive ability. The metrics produced by these baseline models will often depend on the prevalence of each state and/or the number of states. These values can be used as comparators to provide contextual information regarding the strength of model performance. These baselines metrics for the NI-model and the P-model will be referred to as the NI-level and P-level for the metric. In order to allow for simplicity and understanding of these measures, they will be standardised to the same scales.

3.2 Notation

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

[**Notation Table**]

3.3 Patient Weighting

[**Lots of formula, so will leave for now**]