Validation of prediction models in the presence of competing risks: a guide through modern methods

Geloven, N. van; Giardiello, D.; Bonneville, E.F.; Teece, L.; Ramspek, C.L.; Smeden, M. van; ... ; Steyerberg, E.

Supplementary material 4: Technical description of the performance measures

Nan van Geloven et al.

1 General notation

We use the tutorial paper by Putter et al. [1] as main reference for the Sections 1 through 3. We assume that individuals can experience one of K distinct events. We denote the failure time as T, and the competing event indicator as $D \in \{1, ..., K\}$. In practice, individuals are subject to some right-censoring time C, which is assumed to be independent of T and D, possibly given covariates. We thus only observe realisations of $\tilde{T} = \min(C, T)$ and $\tilde{D} = I(T \leq C)D$, where $\tilde{D} = 0$ indicates a right-censored observation and $I(\cdot)$ is the indicator function.

2 Key quantities

The *cause-specific hazard* of failing from a cause k in presence of competing events is defined as:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t}.$$

The overall survival probability is defined by the K cause-specific hazard functions as

$$S(t) = \exp\left(-\sum_{k=1}^{K} \int_0^t h_k(u)du\right) = \exp\left(-\sum_{k=1}^{K} H_k(t)\right),$$

where $H_k(t) = \int_0^t h_k(u)du$ is the cause-specific cumulative hazard for cause k.

The *cumulative incidence function* for an event k, also referred to as the absolute risk of event k, is the probability of that event occurring by a particular time-point t without any other competing event occurring earlier, $P(T \leq t, D = k)$. It is defined as

$$F_k(t) = \int_0^t h_k(u)S(u-)du,$$

with $S(u-)$ being the total survival probability just up to time u.

3 Aalen-Johansen

Suppose we observe n independent samples $(\tilde{t}_i, \tilde{d}_i)$ of $(\tilde{T}, \tilde{D})$, for $i = 1 \cdots n$. We order the J distinct event times where any of the K competing events occur as $0 < t_1 < \ldots < t_J$. Let $D_k(t_j)$ denote the number of individuals failing from cause k at t_j, and let $D(t_j) = \sum_{k=1}^{K} D_k(t_j)$ denote the total number of failures from any cause at t_j. The number of individuals at risk of any event at t_j is given by $R(t_j)$.

The cumulative incidence of cause k by some time horizon s can be estimated non-parametrically using the Aalen-Johansen estimator [2], defined as

$$\widehat{F}_k(s) = \sum_{j:t_j \leq s} \widehat{h}_k(t_j)\widehat{S}(t_{j-1}),$$

where

$$\widehat{h}_k(t_j) = \frac{D_k(t_j)}{R(t_j)}, \qquad \widehat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \sum_{k=1}^{K} \widehat{h}_k(t_j)\right).$$

This Aalen-Johansen estimator is sometimes referred to directly as 'the cumulative incidence function' (e.g. Ramspek et al. [3]). Here we the denote the cumulative incidence function as the population quantity we are targeting, and the Aalen-Johansen estimator as the means to estimate it from data.

4 Regression models

We assume for the remainder of this document that primary interest lies in estimating the cumulative incidence for event $D = 1$ by some prediction horizon s, conditional on covariates. Let $\mathbf{Z}$ denote a vector of p covariates, which are observed for every i^{th} individual as $\mathbf{z}_i$.

The two most commonly used methods for predicting an event conditional on covariates in the presence of competing risks are the Fine and Gray approach [4], and the cause-specific Cox proportional hazards approach. Both are able to produce a subject-specific absolute risk of experiencing event $D = 1$ by s, which we denote as $\pi_1(s \mid \mathbf{z}_i)$. This is effectively an estimate of $F_1(s \mid \mathbf{z}_i) = P(T \leq s, D = 1 \mid \mathbf{z}_i)$.

4.1 Cause-specific Cox proportional hazards approach

The cause-specific approach first entails specifying a Cox proportional hazards model for each of the K competing events as

$$h_k(t \mid \mathbf{Z}) = h_{k0}(t) \exp(\boldsymbol{\beta}_k^{\intercal}\mathbf{Z}),$$

where $h_{k0}(t)$ is the cause-specific baseline hazard, and $\boldsymbol{\beta}_k$ represents the effects of covariates $\mathbf{Z}$ on the cause-specific hazard. Each model can be estimated by treating all events by causes other than $D = k$ as censored. Note that the models need not necessarily share the same covariates.

In order to obtain $\pi_1(s \mid \mathbf{z}_i)$ using the cause-specific approach, the individual-specific hazards must first be calculated as

$$\widehat{h}_k(t \mid \mathbf{z}_i) = \widehat{h}_{k0}(t) \exp(\hat{\boldsymbol{\beta}}_k^{\intercal}\mathbf{z}_i),$$

where $\widehat{h}_{k0}(t)$ is calculated based on the increments in the Breslow estimate of the cause-specific cumulative baseline hazard. These hazards for all J distinct timepoints can thereafter be plugged into the formula for $\widehat{F}_k(s)$ outlined in Section 3, producing $\pi_1(s \mid \mathbf{z}_i)$ for $D = 1$. We refer the reader for example to Section 5.2.1 of the text by Beyersmann et al. [5] for a more detailed treatment of the procedure.

4.2 Fine and Gray approach

The Fine and Gray approach uses a model for the so-called *subdistribution hazard*, defined for cause $D = k$ as

$$\begin{aligned}\lambda_k(t \mid \mathbf{Z}) &= \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t, D = k \mid T \geq t \cup (T \leq t \cap D \neq k), \mathbf{Z}\}}{\Delta t}, \\ &= \frac{-d \log\{1 - F_k(t \mid \mathbf{Z})\}}{dt},\end{aligned}$$

where patients failing from competing causes $D \neq k$ remain in the risk-set up to the end of follow up.

A proportional hazards model can be specified for this subdistribution hazard as

$$\lambda_k(t \mid \mathbf{Z}) = \lambda_{k0}(t) \exp(\boldsymbol{\gamma}_k^\intercal \mathbf{Z}),$$

with $\lambda_{k0}(t)$ being the subdistribution baseline hazard function and $\boldsymbol{\gamma}_k$ representing the effects of covariates $\mathbf{Z}$ on the subdistribution hazard. The cumulative incidence function for $D = k$ can then be written as

$$F_k(s \mid \mathbf{Z}) = 1 - \exp\left[- \exp(\boldsymbol{\gamma}_k^\intercal \mathbf{Z}) \int_0^s \lambda_{k0}(u) du \right],$$

or equivalently,

$$1 - F_k(s \mid \mathbf{Z}) = \{1 - F_{k0}(s)\}^{\exp(\boldsymbol{\gamma}_k^\intercal \mathbf{Z})},$$

where $F_{k0}(s)$ denotes the baseline cumulative incidence. Thus, for event $D = 1$ this model can be used directly to obtain a prediction $\pi_1(s \mid \mathbf{z}_i)$ without having to model the other competing causes.

5 Dealing with censoring when assessing performance

Let T_i and D_i respectively denote the true event time and competing event indicator for an individual i. We can define $Y_i(s) = I(T_i \leq s, D_i = 1)$ as the binary event which indicates whether event $D = 1$ occured prior to the prediction horizon s, or not. If an individual i is censored prior to s, we cannot know whether they would have gone on to experience the event of interest or not. Hence, $Y_i(s)$ is not fully observed in the presence of right-censoring.

5.1 Pseudo-observations

One of the ways to deal with the issue of censoring is to use pseudo-observations $\tilde{Y}_i(s)$ [6], which attempts to recreate $Y_i(s)$. These are defined as

$$\tilde{Y}_i(s) = n\widehat{F}_1(s) - (n-1)\widehat{F}_1^{-i}(s)$$

where $\widehat{F}_1(s)$ is the Aalen-Johansen estimate of $\mathbb{E}\{Y_i(s)\}$ based on all patients, and $\widehat{F}_1^{-i}(s)$ is based on the sample excluding the i^{th} individual. In case of covariate-dependent censoring, a weighted version of the Aalen-Johansen estimator should instead be used [7]. Using $\tilde{Y}_i(s)$ instead of $Y_i(s)$ in the calculation of for instance performance measures eases up calculations as all individuals have a value for $\tilde{Y}_i(s)$.

5.2 IPCW

Another way to deal with the issue of censoring is to use inverse probability of censoring weights (IPCW). Individuals with an observed event status at s are known as a 'complete-case', meaning they have either experienced one of K events prior to s, or are still at risk at s. For those patients experiencing an event at $\tilde{t}_i \leq s$, the probability conditional on covariates $\mathbf{z}_i$ of still being under follow-up just prior to $\tilde{t}_i$ is denoted by $G(\tilde{t}_i- \mid \mathbf{z}_i)$. For those still at risk, $\tilde{t}_i > s$, the probability of being observed to have no event up to time s is written as $G(s \mid \mathbf{z}_i)$. Both can be estimated using the Cox proportional hazards models, or by Kaplan-Meier estimators in absence of any $\mathbf{z}_i$ predictive of censoring.

Individuals who are known to have experienced a particular event before time s or to still be at risk prior to s are then weighted inversely to their probability of having that particular outcome, $1/G(\tilde{t}_i- \mid \mathbf{z}_i)$ or $1/G(s \mid \mathbf{z}_i)$.

6 Performance measures

6.1 Calibration

As per Blanche et al. [8], strong model calibration is defined by

$$\pi_1(s \mid \mathbf{Z}) = P\{Y(s) = 1 \mid \mathbf{Z}\} \qquad \text{for all } \mathbf{Z},$$

meaning that the estimated risk is equal to the observed outcome proportion for all values (and thus combinations) of $\mathbf{Z}$. Unless $\mathbf{Z}$ is low-dimensional and made up entirely of categorical variables, this is typically impossible to assess. We can instead calibration by means of various graphical and numerical summaries.

6.1.1 Calibration plot

The simplest calibration plot bins individuals into approximately equally sized groups based on their risk estimates, and plots the relationship between the average estimated risk and the observed outcome proportion of the event in *each* group. The latter can be either estimated using the Aalen-Johansen estimator, or by averaging across the pseudo-observations within a group. Formally, the calibration plot assesses

$$P\{Y(s) = 1 \mid \pi_1(s \mid \mathbf{Z}) = r\} = r \qquad \text{for all } \mathbf{Z}, \text{ for all } r \in [0, 1],$$

which essentially states that among individuals with an estimated risk of r, the observed outcome proportion should also be r. Methods that attempt to create a *smooth* calibration curve, be it through local smoothing of pseudo-observations [9] or spline-based regression of risk estimates [10], try to create continuity. In other words, they try to make the groups defined by r as small as possible.

Briefly, the *subdistribution model* approach to creating a smooth calibration curve [11] fits a Fine and Gray model for the primary event as a flexible function of the estimated risks, which have been transformed as $\log(-\log(1 - \pi_1(s \mid \mathbf{z}_i)))$. Restricted cubic splines are used as the flexible function, where the number of internal *knots* defines the degree of smoothing. The predictions from this flexible subdistribution model by s serve as the observed outcome proportions, and can be plotted against $\pi_1(s \mid \mathbf{z}_i)$ to create the calibration curve.

The approach taken in [9] to create smooth calibration curves first relies on computing the pseudo-observation $\tilde{Y}_i(s)$ for each individual. Then, for some probability p, the pseudo-observations of individuals with an estimated risk within some interval of p are averaged to

obtain an observed outcome proportion. This pre-specified interval around p, or *bandwidth*, defines the degree of smoothing.

6.1.2 Numerical summaries of calibration

Calibration 'in the large' is defined by

$$\mathbb{E}\{\pi_1(s \mid \mathbf{Z})\} = P\{Y(s) = 1\},$$

stating that the average estimated risk equals the overall observed outcome proportion. A popular way of summarising this is the ratio of cumulative observed over expected events, or O/E. Due to censoring in the current setting, we divide risks instead of absolute event numbers. The observed outcome proportion ('observed') is given by the Aalen-Johansen estimator, while the expected risk is simply the average across all estimated risks. For an alternative calculation of the O/E ratio, see [12].

A second type of numerical summary is the integrated calibration index (ICI), which is a weighted mean of the absolute differences between estimated risks and observed outcome proportions [10]. Specifically, let x represent the vector of estimated risks $\pi_1(s \mid \mathbf{Z})$ by time s, and x_c the value of the calibration curve (i.e. the observed outcome proportions, obtained by smoothing) at x. If we define $f(x) = |x - x_c|$, and define the density function of x as $\phi(x)$, then

$$ICI(s) = \int_0^1 f(x)\phi(x)dx,$$

which is estimated as simply the empirical mean of $f(x)$. The median (E50) and or other percentiles of the $f(x)$ are also possible numerical summaries. Similarly, the squared bias may be of interest, which is estimated as the empirical mean of $f(x) = (x - x_c)^2$.

Note that these numerical summaries depend on the degree and type of smoothing applied to obtain x_c. With higher flexibility, i.e. smaller bandwidth for smoothing the pseudo-observations or higher number of knots in the subdistribution approach, the calibration curve may be overfitted in areas with few observations where the estimated risks are usually very small or large. The advice for the subdistribution approach is to use between 3 and 5 internal knots (Austin et al. 2020+), while for the pseudo-observation approach ample advice is provided in the text by Gerds et al. [9]. Finally, note that the smoothing method chosen to obtain the calibration plot should preferably be the same as the one used when computing the numerical summaries.

A third way to numerically summarize calibration is through the calibration intercept and calibration slope, which additionally allow for miscalibration testing. We briefly explain the application of the methods described in [13] to the competing risks setting. The idea is to model the pseudo-observations $\tilde{Y}_i(s)$ as a function of the complementary log-log transformed estimated risks $\text{cloglog}\{\pi_1(s \mid \mathbf{z}_i)\} = \log(-\log(1 - \pi_1(s \mid \mathbf{z}_i)))$ in a generalized linear regression model (GLM). By writing $\mathbb{E}\{\tilde{Y}(s)\} = \mu$, we can formulate the following two regression models,

$$\text{cloglog}(\mu) = \beta_0 + \text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}, \tag{1}$$

$$\text{cloglog}(\mu) = \beta_0' + \beta_1' \,\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}. \tag{2}$$

Both GLMs use a complementary log-log link function for the mean, and assume constant variance. Additionally, both models are fitted by means of generalized estimating equations (GEE) [14]. Model (1) allows estimation of the calibration intercept β_0, which should ideally be equal to zero. In this model, the transformed risk estimates $\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}$ are used as an *offset*, meaning that its coefficient is constrained to unity. A calibration intercept (significantly) below

or above zero respectively implies on average over and underestimation of the observed outcome proportions.

Model (2) allows estimation of the calibration slope β_1', which should ideally be equal to one. A calibration slope between 0 and 1 indicates too extreme predictions (both on the low and on the high side), while a calibration slope greater than 1 indicates predictions that do not show enough variation. A negative calibration slope implies predictions are in the wrong direction. Furthermore, adding the transformed risk estimates as an offset in model (2) allows to test $\beta_1' = 1$ directly.

Regarding testing, it is preferable to first perform a joint test $(\beta_0', \beta_1') = (0, 1)$ with two degrees of freedom to assess overall evidence for miscalibration [15]. If the null-hypothesis is rejected in the joint test, the individual tests for β_0 and β_1' can then be performed.

6.2 Discrimination

We introduce a pair of individuals i and j with covariates $\mathbf{z}_i$ and $\mathbf{z}_j$ respectively. At horizon s, we have model-based predictions $\pi_1(s \mid \mathbf{z}_i)$ and $\pi_1(s \mid \mathbf{z}_j)$. The ordering of these estimated risks at s is thus denoted by

$$Q_{ij}(s) = I\big\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j)\big\}.$$

6.2.1 C-index

As described in [16], the 'truncated' concordance index (C-index) is defined by

$$\mathcal{C}_1(s) = P\big\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j) \mid D_i = 1, T_i \leq s, (T_i < T_j \cup D_j \notin \{0,1\})\big\}.$$

It measures how well the model ranks the event times occurring prior to s [17]. Notice that for a pair of individuals, if the individual with the earlier event time is right-censored, the ordering $T_i < T_j$ is indeterminable. A simple solution for estimating the C-index is setting the follow up time of the patients with competing event to the maximum follow up time in the study design [18]. This method can however only be used in settings without censoring or with purely administrative censoring, as recently illustrated for prediction of kidney failure [19]. To estimate the C-index in the presence of other types of right-censoring, we can construct weights as part of an IPCW procedure, yielding

$$w_{ij,1} = \frac{I(\tilde{t}_i < \tilde{t}_j)}{\widehat{G}(\tilde{t}_i- \mid \mathbf{z}_i)\widehat{G}(\tilde{t}_i \mid \mathbf{z}_j)}, \qquad w_{ij,2} = \frac{I(\tilde{t}_i \geq \tilde{t}_j, \tilde{d}_j \notin \{0,1\})}{\widehat{G}(\tilde{t}_i- \mid \mathbf{z}_i)\widehat{G}(\tilde{t}_j- \mid \mathbf{z}_j)}.$$

We can then estimate the c-index as

$$\widehat{\mathcal{C}}_1(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n (w_{ij,1} + w_{ij,2}) Q_{ij}(s) I(\tilde{t}_i \leq s, \tilde{d}_i = 1)}{\sum_{i=1}^n \sum_{j=1}^n (w_{ij,1} + w_{ij,2}) I(\tilde{t}_i \leq s, \tilde{d}_i = 1)}.$$

We note that the c-index is not appropriate for validating prediction models with time-varying covariate effects [20].

6.2.2 Time-dependent area under the ROC curve

We define cases as individuals with $\tilde{t}_i \leq s$ and $\tilde{d}_i = 1$, i.e. as experiencing the primary event by s. Controls however have been defined in two ways:

1. free of any event by s, i.e. $\tilde{t}_i > s$,
2. free of any event by s, i.e. $\tilde{t}_i > s$, or experiencing a competing event, $(\tilde{t}_i \leq s,\, \tilde{d}_i \notin \{0,1\})$.

We continue with the second definition here. We define a time-dependent area under the receiving operating characteristic curve (AUC_t), described in [21] and the supplementary material of [16]. It is defined as

$$AUC_1(s) = P\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j) \mid D_i = 1, T_i \leq s, (T_j > s \cup D_j \notin \{0,1\})\}.$$

It evaluates the concordance of risk estimates between individuals experiencing the primary event by s, and individuals either event-free or that have experienced a competing event. Similarly to the C-index, a pair becomes unevaluable (directly) if one of the individuals has a right-censored event time prior to s. Specifically, we cannot determine whether this individual would experience the primary event between the right-censoring time and s, or remain a control. Thus, we must first construct weights

$$w_i = \frac{I(\tilde{t}_i \leq s, \tilde{d}_i = 1)}{\hat{G}(\tilde{t}_i)}, \qquad w_{j,1} = \frac{I(\tilde{t}_j \leq s, \tilde{d}_j \notin \{0,1\})}{\hat{G}(\tilde{t}_j)}, \qquad w_{j,2} = \frac{I(\tilde{t}_j > s)}{\hat{G}(s)},$$

and then can estimate $AUC_1(s)$ as

$$\widehat{AUC}_1(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_i(w_{j,1} + w_{j,2})Q_{ij}(s)}{\sum_{i=1}^n w_i \sum_{j=1}^n (w_{j,1} + w_{j,2})}.$$

We refer to [21] for details on covariate dependent censoring.

Alternative versions of the AUC_t have been proposed which use different definitions of cases and controls according to having their events before, at or after the time-point of interest [22]. The cumulative case/dynamic control definition we describe here can be considered most suited for evaluation of predictions from baseline over a specific prediction horizon [23] whereas the incident case/dynamic control definition with cases defined as having the primary event exactly at (i.e. not before) a fixed time-point, can be useful in evaluating dynamic prediction models [23, 24, 25].

6.3 Overall prediction error

6.3.1 Brier score

The Brier score in the context of competing events is the expected quadratic distance between the event indicator $Y(s)$ (for the primary event $D = 1$) and the estimated risks $\pi_1(s \mid \mathbf{Z})$ based on the prediction model,

$$B_1(s) = \mathbb{E}\big[I(T \leq s, D = 1) - \pi_1(s \mid \mathbf{Z})\big]^2,$$

with $I(T \leq s, D = 1)$ being the true event status at s. In the presence of censoring, the Brier score can be estimated using either IPCW, or pseudo-observations. The latter estimator has been suggested in the context of dynamic prediction [26], and in the context of multistate models in [27].

As per Schoop et al. [28], an IPCW estimator for the Brier score is

$$\widehat{B}_1(s) = \frac{1}{n}\sum_{i=1}^n \big[I(\tilde{t}_i \leq s, \tilde{d}_i = 1) - \pi_1(s \mid \mathbf{z}_i)\big]^2 w_{1i},$$

where

$$w_{1i} = \frac{I(\tilde{t}_i \leq s, \tilde{d}_i \neq 0)}{\widehat{G}(\tilde{t}_i - \mid \mathbf{z}_i)} + \frac{I(\tilde{t}_i > s)}{\widehat{G}(s \mid \mathbf{z}_i)}.$$

6.3.2 Scaled Brier score

As per Kattan and Gerds [29], the scaled Brier score (also know as index of prediction accuracy, IPA) for estimating the cumulative incidence of event $D = 1$ is

$$\text{IPA}(s) = 1 - \frac{B_1^{\text{mod}}(s)}{B_1^{\text{null}}(s)},$$

where $B_1^{\text{mod}}(s)$ is the model Brier score, and $B_1^{\text{null}}(s)$ is the Brier score for the null model (with no covariates). The latter can be calculated by plugging-in the Aalen-Johansen estimator in place of $\pi_1(s \mid \mathbf{z}_i)$.

6.4 Decision curve analysis

In a competing risks setting, the net benefit at s based on a prediction model for the primary event, given a chosen probability threshold p_s, is given by

$$\text{NB}_1(s) = \frac{\text{TP}_1(s)}{n} - \frac{\text{FP}_1(s)}{n}\left(\frac{p_s}{1-p_s}\right), \tag{3}$$

where $\text{TP}_1(s)$ is the true positive count and $\text{FP}_1(s)$ the false positive count.

In order to estimate the net benefit, we first define $x_i = 1$ if $\pi_1(s \mid \mathbf{z}_i) \geq p_s$. In other words, x_i defines whether an individual is classified as their estimated risk exceeding the chosen probability threshold p_s. $P(X = 1)$ is then the proportion classified as $X = 1$ based on this threshold.

Recall $\widehat{F}_1(s)$ as the Aalen-Johansen estimate of the cumulative incidence of event $D = 1$ by horizon s. The quantity $\widehat{F}_1(s \mid X = 1)$ is then the estimated cumulative incidence *among* those classified as exceeding the risk threshold. As described in [30], the number of true positives is estimated as

$$\widehat{\text{TP}}_1(s) = \widehat{F}_1(s \mid X = 1) \times P(X = 1) \times n,$$

and similarly the number of false positives as

$$\widehat{\text{FP}}_1(s) = \left[1 - \widehat{F}_1(s \mid X = 1)\right] \times P(X = 1) \times n.$$

The estimated net-benefit $\widehat{\text{NB}}_1(s)$ can then be obtained by plugging-in $\widehat{\text{TP}}_1(s)$ and $\widehat{\text{FP}}_1(s)$ into (3). Furthermore, a *decision curve* can be obtained by plotting $\widehat{\text{NB}}_1(s)$ for various values of p_s. This is often plotted alongside a 'treat-all' curve, which plots the net-benefit across thresholds in a situation where all individuals are classified as exceeding the risk threshold regardless of the prediction model. A 'treat none' reference line is useful as well, with net-benefit of zero for any threshold (no true positive and no false positive decisions are made).

7 Closing remarks

Formulas concerning standard errors of performance measures are beyond the scope of this document. Analytical formulas are available for many measures, and bootstrapping can be used for most.

References

[1] H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. ISSN 1097-0258.

[2] Odd O. Aalen and Søren Johansen. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3): 141–150, 1978. ISSN 0303-6898.

[3] Chava L Ramspek, Lucy Teece, Kym I E Snell, Marie Evans, Richard D Riley, Maarten van Smeden, Nan van Geloven, and Merel van Diepen. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. *International Journal of Epidemiology*, page dyab256, December 2021. ISSN 0300-5771, 1464-3685.

[4] Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. ISSN 0162-1459.

[5] Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing Risks and Multistate Models with R.* Use R! Springer-Verlag, New York, 2012. ISBN 978-1-4614-2034-7.

[6] Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis:. *Statistical Methods in Medical Research*, August 2009.

[7] Nadine Binder, Thomas A. Gerds, and Per Kragh Andersen. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, 20(2):303–315, April 2014. ISSN 1572-9249.

[8] Paul Blanche, Thomas A. Gerds, and Claus T. Ekstrøm. The Wally plot approach to assess the calibration of clinical prediction models. *Lifetime Data Analysis*, 25(1):150–167, January 2019. ISSN 1572-9249.

[9] Thomas A. Gerds, Per K. Andersen, and Michael W. Kattan. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191–3203, 2014. ISSN 1097-0258.

[10] Peter C. Austin, Frank E. Harrell, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Statistics in Medicine*, 39 (21):2714–2742, 2020. ISSN 1097-0258.

[11] Peter C. Austin, Hein Putter, Daniele Giardiello, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. *Diagnostic and Prognostic Research*, 6(1):2, January 2022. ISSN 2397-7523.

[12] Adam R. Brentnall and Jack Cuzick. Risk Models for Breast Cancer and Their Validation. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 35(1):14–30, March 2020. ISSN 0883-4237.

[13] Patrick Royston. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. *The Stata Journal*, 14(4):738–755, December 2014. ISSN 1536-867X.

[14] Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics. Journal of the International Biometric Society*, pages 121–130, 1986.

[15] D. R. Cox. Two Further Applications of a Model for Binary Regression. *Biometrika*, 45 (3/4):562–565, 1958. ISSN 0006-3444.

[16] Marcel Wolbers, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, July 2014. ISSN 1465-4644.

[17] Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, June 2013. ISSN 02776715.

[18] Marcel Wolbers, Michael T. Koller, Jacqueline C. M. Witteman, and Ewout W. Steyerberg. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology*, 20(4):555–561, July 2009. ISSN 1044-3983.

[19] Chava L. Ramspek, Marie Evans, Christoph Wanner, Christiane Drechsler, Nicholas C. Chesnaye, Maciej Szymczak, Magdalena Krajewska, Claudia Torino, Gaetana Porto, Samantha Hayward, Fergus Caskey, Friedo W. Dekker, Kitty J. Jager, Merel van Diepen, and the EQUAL Study Investigators. Kidney Failure Prediction Models: A Comprehensive External Validation Study in Patients with Advanced CKD. *Journal of the American Society of Nephrology*, 32(5):1174–1186, May 2021. ISSN 1046-6673, 1533-3450.

[20] Janez Stare, Maja Pohar Perme, and Robin Henderson. A Measure of Explained Variation for Event History Data. *Biometrics*, 67(3):750–759, 2011. ISSN 1541-0420.

[21] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, December 2013. ISSN 02776715.

[22] Patrick J. Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, March 2005. ISSN 0006-341X, 1541-0420.

[23] P. Saha and P. J. Heagerty. Time-Dependent Predictive Accuracy in the Presence of Competing Risks. *Biometrics*, 66(4):999–1011, December 2010. ISSN 0006341X.

[24] Hans van Houwelingen and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, zeroth edition, November 2011. ISBN 978-0-429-09433-0.

[25] N. van Geloven, Y. He, A. H. Zwinderman, and H. Putter. Estimation of incident dynamic AUC in practice. *Computational Statistics & Data Analysis*, 154:107095, February 2021. ISSN 0167-9473.

[26] Giuliana Cortese, Thomas A. Gerds, and Per K. Andersen. Comparing predictions among competing risks models with time-dependent covariates. *Statistics in Medicine*, 32(18): 3089–3101, 2013. ISSN 1097-0258.

[27] Cristian Spitoni, Violette Lammens, and Hein Putter. Prediction errors for state occupation and transition probabilities in multi-state models. *Biometrical Journal*, 60(1):34–48, 2018. ISSN 1521-4036.

[28] Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, February 2011. ISSN 03233847.

[29] Michael W. Kattan and Thomas A. Gerds. The index of prediction accuracy: An intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research*, 2(1):7, December 2018. ISSN 2397-7523.

[30] Andrew J Vickers, Angel M Cronin, Elena B Elkin, and Mithat Gonen. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*, 8(1):53, December 2008. ISSN 1472-6947.