Assessing the calibration of subdistribution hazard models in discrete time

Moritz BERGER1* and Matthias SCHMID1

1Institute of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, Germany

Key words and phrases: Calibration; competing risks; discrete time-to-event data; subdistribution hazard; validation

MSC 2020: Primary 62N01; secondary 62P10.

Abstract: The generalization performance of a risk prediction model can be evaluated by its calibration, which measures the agreement between predicted and observed outcomes on external validation data. Here, we propose methods for assessing the calibration of discrete time-to-event models in the presence of competing risks. Specifically, we consider the class of discrete subdistribution hazard models, which directly relate the cumulative incidence function of one event of interest to a set of covariates. We apply the methods to a prediction model for the development of nosocomial pneumonia. Simulation studies show that the methods are strong tools for calibration assessment even in scenarios with a high censoring rate and/or a large number of discrete time points. The Canadian Journal of Statistics 50: 572–591; 2022 © 2021 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada.

Rédsumé: La performance de généralisation d’un modèle de prévision des risques peut être évaluée par sa calibration qui mesure la concordance entre les valeurs prédites et observées dans des données externes de validation. Les auteurs proposent des méthodes pour évaluer la calibration de modèles discrets de durée de vie en présence de risques concurrents. Plus précisément, ils considèrent la classe de modèles à sous-distribution discrète du risque qui relie directement la fonction d’incidence cumulative d’un événement à un ensemble de covariables. Les auteurs appliquent leurs méthodes à un modèle de prévision pour le développement de pneumonie nosocomiale. Ils présentent des études de simulation montrant que les méthodes sont d’excellents outils pour l’évaluation de la calibration, même dans les scénarios comportant un haut taux de censure et/ou un large nombre de points temporels discrets. La revue canadienne de statistique 50: 572–591; 2022 © 2021 Les auteurs. La revue canadienne de statistique/The Canadian Journal of Statistics, publiée par Wiley Periodicals LLC au nom de la Société statistique du Canada.

1. INTRODUCTION

Over the past decade, risk prediction models have become an indispensable tool for decision making in applied research. Popular examples include models for diagnosis and prognosis in the health sciences where risk prediction is used, such as screening and therapy decisions (Moons et al., 2012b; Liu et al., 2014; Steyerberg, 2019) and models for risk assessment in ecological research, which have become an established tool to quantify and forecast the ecological impact of technology and development (Gibbs, 2011).

Additional Supporting Information may be found in the online version of this article at the publisher’s website.

* Corresponding author: moritz.berger@imbie.uni-bonn.de

© 2021 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
A key aspect in the development of risk prediction models is the validation of generalization performance. This task, which is usually performed by applying a previously derived candidate prediction model to one or more sets of independent external validation data, has been subject to extensive methodological research (Moons et al., 2012a; Steyerberg & Vergouwe, 2014; Harrell, 2015; Steyerberg & Harrell, 2016; Alba et al., 2017). As a result, strategies for investigating the discriminatory power (measuring how well a model separates cases from controls), calibration (measuring the agreement between predicted and observed outcomes) and prediction error (quantifying both discrimination and calibration aspects) of prediction models have been developed (Steyerberg et al., 2010). Alternative techniques that additionally involve decision analytic measures include, among others, net benefit analysis (Vickers, Van Calster & Steyerberg, 2016), decision curve analysis (Vickers & Elkin, 2006) and relative utility curve analysis (Baker et al., 2009; Kerr & Janes, 2017).

The aim of this article is to develop a set of methods for assessing the calibration of a prediction model with a time-to-event outcome. This class of models has been dealt with extensively during the past years, see, for example, Henderson & Keiding (2005), Witten & Tibshirani (2010), Soave & Strug (2018) and Braun et al. (2018). Here, we explicitly assume that event times are measured on a discrete time scale \( t = 1, 2, \ldots \) (Ding et al., 2012; Tutz & Schmid, 2016; Berger & Schmid, 2018), and that the event of interest may occur along with one or more “competing” events (Fahrmeir & Wagenpfeil, 1996; Fine & Gray, 1999; Lau, Cole & Gange, 2009; Beyersmann, Allignol & Schumacher, 2011; Austin, Lee & Fine, 2016; Lee, Feuer & Fine, 2018; Schmid & Berger, 2020). Scenarios of this type are frequently encountered in observational studies with a limited number of fixed follow-up measurements, for instance, in epidemiology (Andersen et al., 2012). Such study designs do not allow recording the exact (continuous) event times, so that it is only known whether or not an event of interest (or a competing event) occurred between two consecutive follow-up times \( a_{r-1} \) and \( a_r \), implying that the discrete time scale \( t = 1, 2, \ldots \) refers to a special case of interval censoring with fixed intervals.

An important example, which will be considered in this article, is the duration to the development of nosocomial pneumonia (NP) in intensive care patients measured on a daily basis (Wolkewitz et al., 2008). As NP infections are associated with an increased length of hospital stay and have considerable impact on morbidity and mortality, it is highly relevant to build a statistical model that gives valid predictions for future patients. The case of observational hospital data is interesting in that early discrete-time competing risks analysis is found in the literature as early as the 1860s (Nightingale, 1863, Chapter IX). In Section 7 we validate a prediction model developed by Berger et al. (2020) for this type of data.

In recent years, several authors have developed measures and estimators for analyzing the generalization performance of discrete time-to-event models. For example, discrimination measures for discrete time-to-event models were proposed by Schmid, Tutz & Welchowski (2018). Measures of prediction error were considered in Tutz & Schmid (2016), Chapter 4. Graphical tools for assessing the calibration of discrete time-to-event predictions (not accounting for the occurrence of competing events) were explored in Berger & Schmid (2018). Methods for assessing the generalization performance of discrete cause-specific hazard models (a common approach for competing risks analysis) have been recently proposed by Heyard et al. (2020).

Here we propose to base the calibration assessments for discrete competing risks models on the cumulative incidence function \( F_j(t|x) := P(T \leq t, e = j|x) \), denoting by \( T \) the time to the first event, by \( x \) a set of covariates, and by \( e \in \{1, \ldots, J\} \) a random variable that indicates the occurrence of one out of \( J \) competing events at \( T \) (Fine & Gray, 1999; Klein & Andersen, 2005). In the following, we will assume without loss of generality that the event of interest and its cumulative incidence function are defined by \( e = 1 \) and \( F_1(t|x) \), respectively. A popular method to derive
predictions of $F_1(t|x)$ from a set of training data is to fit a proportional subdistribution hazard model (Fine & Gray, 1999). This approach is designed for the analysis of right-censored event times, and it will be considered here. This approach has been recommended to analysts “whenever the focus is on estimating incidence or predicting prognosis in the presence of competing risks” (Austin, Lee & Fine, 2016). While the original model proposed by Fine & Gray (1999) assumed the event times to be measured on a continuous time scale, the methods developed in this article are based on a recent extension of the subdistribution hazard modelling approach to discrete-time competing risks data (Berger et al., 2020). Specifically, the model proposed by Berger et al. (2020) is designed for estimating the subdistribution hazards along the lines of Cox (1958) and Miller et al. (1993). As will be shown in Section 2 and 3 for details). The calibration of a subdistribution hazard prediction model may thus be characterized by how well the subdistribution hazards observed in a validation sample can be approximated by the respective predicted subdistribution hazards that are obtained from applying the prediction model to the validation data.

The proposed methodology for assessing the calibration of a discrete-time subdistribution hazard model comprises two parts, both of which build on methods for binary regression: The first part (presented in Section 4) will be concerned with the derivation of an appropriate calibration plot that visualizes the agreement between the predicted and the observed subdistribution hazards. In the second part (Section 5), we will propose a recalibration model for discrete-time subdistribution hazard models that can be used to analyze calibration-in-the-large and refinement (i.e., the bias and the variation, respectively, of the predicted subdistribution hazards along the lines of Cox (1958) and Miller et al. (1993). As will be shown in Sections 4 and 5, the weights used in the subdistribution hazard modelling approach proposed in Berger et al. (2020) allow defining appropriate versions of the observed and predicted hazards (to be depicted in the calibration plot) and for fitting a weighted logistic recalibration model (giving rise to point estimates and hypothesis tests on calibration-in-the-large and refinement).

The proposed calibration assessments will be illustrated by a simulation study (Section 6) and by the aforementioned prediction model for the duration to the development of NP (Section 7). Section 8 summarizes the main findings of the article.

2. DISCRETE SUBDISTRIBUTION HAZARD MODELS

Let $T_i$ be the event time and $C_i$ be the censoring time of an i.i.d. sample with $n$ individuals $i = 1, \ldots, n$. Both $T_i$ and $C_i$ are assumed to be independent random variables (random censoring) taking discrete values in $\{1, 2, \ldots, k\}$, where $k$ is a natural number. It is further assumed that the censoring mechanism is non-informative for $T_i$, in the sense that $C_i$ does not depend on any parameters used to model the event time (Kalbfleisch & Prentice, 2002). For instance, in longitudinal studies with fixed follow-up visits, the discrete event times $1, \ldots, k$ may refer to time intervals $[0, a_1), [a_1, a_2), \ldots, [a_{k-1}, \infty)$, where $T_i = t$ means that the event has occurred in time interval $[a_{t-1}, a_t)$ with $a_k = \infty$. For right-censored data, the time period during which an individual is under observation is denoted by $\hat{T}_i = \min(T_i, C_i)$, that is, $\hat{T}_i$ corresponds to the true event time if $T_i \leq C_i$ and to the censoring time otherwise. The random variable $\Delta_i := I(T_i \leq C_i)$ indicates whether $\hat{T}_i$ is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). Here, it is assumed that each individual can experience one out of $J$ competing events and that the event type of the $i$th individual at $T_i$ is denoted by $e_i \in \{1, \ldots, J\}$. In accordance with Fine & Gray (1999), our interest is in modelling the cumulative incidence function $F_1(t) = P(T \leq t, e = 1)$ of a type 1 event conditional on covariates, taking into account that there are $J - 1$ competing events and also the censoring event ($\Delta_i = 0$).
For given values of a set of time-constant covariates $x_i = (x_{i1}, \ldots, x_{ip})^T$, the discrete subdistribution hazard function for a type 1 event is defined by

$$
\lambda_1(t|x_i) = P \left( T_i = t, \epsilon_i = 1 | (T_i \geq t) \cup (T_i \leq t - 1, \epsilon_i \neq 1), x_i \right)
$$

(1)

$$
P(\theta_i = t | \theta_i \geq t) = P(\theta_i = t | \theta_i \leq t - 1, \epsilon_i \neq 1, x_i)
$$

(2)

where (2) is the discrete hazard function of the subdistribution time

$$
\theta_i := \begin{cases} T_i, & \text{if } \epsilon_i = 1, \\ \infty, & \text{if } \epsilon_i \neq 1, \end{cases}
$$

(3)

see Berger et al. (2020). The subdistribution time $\theta_i$ measures the time to the occurrence of a type 1 event first and is not finite if $\epsilon_i \neq 1$ (as a type 1 event will never be the first event as soon as a competing event has occurred). The discrete subdistribution hazard is linked to the cumulative incidence function by

$$
F_1(t|x_i) = 1 - \prod_{s=1}^k (1 - \lambda_1(s|x_i)) = 1 - S_1(t|x_i),
$$

(4)

where $S_1(t|x_i) = P(\theta_i > t | x_i)$ is the discrete survival function for a type 1 event. Thus, a regression model for the discrete subdistribution hazard $\lambda_1$ has a direct interpretation in terms of the cumulative incidence function $F_1$.

A class of regression models that relate the discrete subdistribution hazard function (2) to the covariates $x_i$ was proposed by Berger et al. (2020). It is defined by

$$
\lambda_1(t|x_i) = h(\eta_1(t, x_i)),
$$

(5)

where $h(\cdot)$ is a strictly monotone increasing distribution function. In line with classical hazard models for discrete event times (e.g., Tutz & Schmid 2016), it is assumed that the predictor function

$$
\eta_1(t, x_i) = \gamma_0 \tau + x_i^T \gamma
$$

(6)

is composed of a set of time-varying intercepts $\gamma_{01}, \ldots, \gamma_{0,k-1}$, referred to as baseline coefficients, and a linear function of the covariates with coefficients $\gamma \in \mathbb{R}^p$ that do not depend on $t$. As in generalized additive models, it is also possible to extend $\eta_1(t, x_i)$ by interactions and smooth (possibly nonlinear) functions. A popular choice of $h(\cdot)$ is the inverse complementary log–log function, which yields the Gompertz model $\lambda_1(t, x_i) = 1 - \exp(-\exp(\eta_1(t, x_i)))$, which is equivalent to the original subdistribution hazard model by Fine & Gray (1999) for continuous time-to-event data.

3. MODEL FITTING

In Berger et al. (2020), it was shown that consistent estimates of the model parameters in (6) can be derived using estimation techniques for weighted binary regression. This result is based on the observation that with i.i.d. data $(\tilde{T}_i, \Delta_i, \epsilon_i, x_i)$, $i = 1, \ldots, n$, the log likelihood of model (5) can be expressed as

$$
\ell = \sum_{i=1}^n \sum_{t=1}^{k-1} w_{it} \left\{ y_{it} \log(\lambda_1(t|x_i)) + (1 - y_{it}) \log(1 - \lambda_1(t|x_i)) \right\}
$$

(7)
with binary outcome values
\[
(y_{i1}, \ldots, y_{i\tilde{T}_i}, \ldots, y_{i,k-1}) = \begin{cases} 
(0, \ldots, 0, 1, 0, \ldots, 0), & \text{if } \Delta_i \varepsilon_i = 1, \\
(0, \ldots, 0, 0, 0, \ldots, 0), & \text{if } \Delta_i \varepsilon_i \neq 1.
\end{cases}
\] (8)

For uncensored individuals that experience a type 1 event first \((\Delta_i \varepsilon_i = 1)\) and for censored individuals \((\Delta_i \varepsilon_i = 0)\) the weights \(w_{it}\) are defined as
\[
w_{it} := 1(t \leq \tilde{T}_i),
\] whereas for uncensored individuals experiencing a competing event first \((\Delta_i \varepsilon_i > 1)\) they are defined as
\[
w_{it} := \begin{cases} 
1 & \text{if } t \leq \tilde{T}_i, \\
\frac{\hat{V}(t-1)}{\hat{V}(\tilde{T}_i-1)} & \text{if } \tilde{T}_i < t \leq k - 1.
\end{cases}
\] (10)

The function \(\hat{V}(t)\) in (10) is an estimate of the censoring survival function \(V(t) = P(C_j > t)\), implying that the weights in (9) and (10) equal estimates of the individual-specific conditional probabilities of being (still) at risk for a type 1 event at time \(t\). This is analogous to the respective weights for subdistribution hazard models in continuous time (Fine & Gray, 1999). As shown in Berger et al. (2020), maximization of the log likelihood (7) yields consistent and asymptotically normal estimators of the parameters \(\gamma_0\) and \(\gamma\). In Sections 4 and 5, we will show that the weights defined in (9) and (10) also play a key role in the calibration assessment of discrete-time subdistribution hazard models.

It should be noted that the inclusion of time-varying covariates in model (6) is not without problems. This is because the weighted log likelihood in (7) requires the time-dependent values of \(x_i\) to be known up to time point \(k - 1\) and thus possibly beyond the observed event times \(\tilde{T}_i\). When the covariates are not external, this is often unrealistic or even impossible. In particular, the cumulative incidence function (4) cannot be written as a function of the hazards when the model includes random (internal) time-varying covariates (Cortese & Andersen, 2010). Finding an adequate strategy for the analysis of such covariates in the subdistribution hazard modelling framework remains challenging (Poguntke et al., 2018; Schmid & Berger, 2020).

4. CALIBRATION PLOT

In the following we will assume that an i.i.d. training sample \((\tilde{T}_i, \Delta_i, \varepsilon_i, x_i), i = 1, \ldots, n\) has been used to fit a statistical model that can be used to predict the individual-specific subdistribution hazards \(\lambda_i(t|x)\) in some study population. We will further assume that the calibration of the fitted model is assessed by means of an independent i.i.d. validation sample with \(N\) individuals \((\tilde{T}_m, \Delta_m, \varepsilon_m, x_m), m = 1, \ldots, N\). The starting point of our considerations is the calibration plot proposed in Berger & Schmid (2018), which applies to discrete hazard models with only a single type of event \((J = 1)\). Note that both the specification of the subdistribution hazard model and the definition of its log-likelihood function remain valid in this case, as the scenario without competing events \((J = 1)\) is a special case of Equations (5) and (7). The idea underlying the method by Berger & Schmid (2018) is to split the test data into \(G\) subsets \(D_g, g = 1, \ldots, G\), defined by the percentiles of the predicted hazards \(\hat{\lambda}_i(t|x_m) = \hat{P}(T_m = t | T_m \geq t, x_m) = \hat{P}(y_{mt} = 1| x_m), t = 1, \ldots, \tilde{T}_m, m = 1, \ldots, N\), which are obtained from the fitted binary model in (5). Following the approach by Hosmer, Lemeshow & Sturdivant (2013) for assessing the calibration of binary regression models, the average predicted hazards in the \(G\) groups are subsequently plotted against the empirical
hazards, which are given by the group-wise relative frequencies of outcome values with \( y_{mt} = 1 \). A well-calibrated model is indicated by a set of points that is close to the 45-degree line.

More formally, the predicted and empirical hazard estimates considered in Berger & Schmid (2018) can be written as

\[
\hat{\lambda}_{1g} = \frac{1}{\sum_{m,t: \hat{\lambda}_1(t|x_m) \in D_g} w_{mt}} \sum_{m,t: \hat{\lambda}_1(t|x_m) \in D_g} \hat{\lambda}_1(t|x_m) w_{mt},
\]

and

\[
\bar{y}_g = \frac{1}{\sum_{m,t: \hat{\lambda}_1(t|x_m) \in D_g} w_{mt}} \sum_{m,t: \hat{\lambda}_1(t|x_m) \in D_g} y_{mt} w_{mt},
\]

\( t = 1, \ldots, k - 1, m = 1, \ldots, N, \) \hspace{1cm} \hspace{1cm} \hspace{1cm} (11)

respectively, where \( w_{mt} := I(t \leq \tilde{T}_m) \in \{0, 1\} \) indicates whether individual \( m \) is at risk for a type 1 event at time point \( t \), \( t = 1, \ldots, k - 1 \), or not. Note that the definition of \( w_{mt} \) in (11) is exactly the same as the definition of the weight \( w_i \) in (9). Also note that \( \sum_{m,t: \hat{\lambda}_1(t|x_m) \in D_g} w_{mt} = |D_g| \), as only the values \( \hat{\lambda}_1(t|x_m) \) with \( w_{mt} = 1 \) are used for defining the groups \( D_1, \ldots, D_g \). In a well-calibrated hazard model, the values \( \bar{\lambda}_{1g} \) should be close to their counterparts \( \bar{y}_g \).

Now consider the scenario where, in addition to the type 1 event of interest, competing events of type \( 2, \ldots, J \) may be observed. In this case, \( \hat{\lambda}_1 \) becomes the subdistribution hazard of a type 1 event, as defined in (2). To obtain a calibration plot for a fitted subdistribution hazard model, we define the quantities \( \bar{\lambda}_{1g} \) and \( \bar{y}_g \) analogous to the single-event scenario considered in (11). Unlike in the scenario with \( J = 1 \), however, the definition of the terms \( w_{mt} \) is not straightforward: the problem is that individuals experiencing a competing event first continue to be at risk beyond \( \tilde{T}_m \) until they experience the censoring event. Hence, as the censoring times \( C_m \) are unobserved if \( C_m > \tilde{T}_m \), it usually cannot be determined whether these individuals would still be at risk at \( t > \tilde{T}_m \). In accordance with Berger et al. (2020), we therefore propose to predict the probability of each individual \( m = 1, \ldots, N \), of being at risk for a type 1 event at time \( t \) and to set the terms \( w_{mt} \) equal to the predicted probabilities.

More specifically, the proposed strategy comprises the following steps:

(i) Sort the predicted subdistribution hazards \( \hat{\lambda}(t|x_m), t = 1, \ldots, k - 1, m = 1, \ldots, N \), obtained from the fitted subdistribution hazard model and form groups \( D_1, \ldots, D_G \) defined by the percentiles of \( \hat{\lambda}(t|x_m) \).

(ii) Compute the weights \( w_{mt} \) using the formulas in (9) and (10), where \( \hat{V}(\cdot) \) is estimated from the training sample with individuals \( i = 1, \ldots, n \).

(iii) Compute \( \bar{\lambda}_{1g} \) and \( \bar{y}_g \) as in (11) using the weights obtained in step (ii). Note that by definition, \( \sum_{m,t: \hat{\lambda}_1(t|x_m) \in D_g} w_{mt} \leq |D_g| \).

(iv) Plot \( \bar{\lambda}_{1g} \) against \( \bar{y}_g \) (using proportional axes).

(v) Assess the calibration of the fitted subdistribution hazard model by inspecting the plot generated in step (iv). A well-calibrated model is indicated by a set of points that is close to the 45-degree line.

For the choice of \( G \), we propose to use the rule by Doane (1976), which was originally developed for univariate frequency classification. With this rule, the number of subsets is defined by

\[
G = \left[ 1 + \log_2(\mathcal{N}) + \log_2\left(1 + \frac{\kappa_{\hat{\lambda}_1}}{\sigma_{\hat{\lambda}_1}}\right) \right],
\]

\hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} (12)
where $\kappa_{\lambda_1}$ denotes the skewness of the predicted subdistribution hazards, $N' = N \cdot (k - 1)$, and $\sigma_{\kappa_{\lambda_1}} = \sqrt{6 (N' - 2)/(N' + 1)/(N' + 3)}$.

It should be emphasized that the calibration plot constitutes an exploratory approach, and that inspection of the plot in step (v) generally involves subjective impression. Formal tests on calibration are proposed in the next section.

5. RECALIBRATION MODEL

In addition to the graphical checks presented in Section 4, we propose a recalibration approach for discrete subdistribution hazard models originating from the method by Cox (1958). The idea of this method, which was originally developed for assessing the calibration of binary regression models, is to fit a logistic regression model to the test data in order to investigate the agreement between a set of predicted probabilities and the respective values of the binary outcome variable.

Based on the binary representation of the subdistribution hazard model in (5) and (7), we propose to adapt the recalibration framework by Cox (1958) as follows: assuming that calibration assessments are again based on a validation sample $(T_m, \Delta_m, \epsilon_m, x_m)$, $m = 1, \ldots, N$, we fit a logistic regression model of the form

$$
\log \left( \frac{\hat{\lambda}_1(t|x_m)}{1 - \hat{\lambda}_1(t|x_m)} \right) = \eta_{tc}(t|x_m) = a + b \log \left( \frac{\hat{\lambda}_1(t|x_m)}{1 - \hat{\lambda}_1(t|x_m)} \right),
$$

for $t = 1, \ldots, k - 1$, $m = 1, \ldots, N,

(13)

where $\hat{\lambda}_1(t|x_m)$ are the predicted hazards defined in Section 4.

In (13) a simple linear model is placed on the logits of the subdistribution hazards. Alternatively, one could also use other link functions, like the probit link or complementary log–log link. The intercept $a$ in model (13) measures “calibration-in-the-large,” that is, it indicates whether the predicted hazards are systematically too low ($a > 0$) or too high ($a < 0$). Analogously, the slope $b$ measures “refinement,” which indicates that the predicted hazards do not show enough variation ($b > 1$), show too much variation ($0 < b < 1$), or show the wrong general direction ($b < 0$, Miller et al., 1993).

To assess the fit of the predicted hazards, we propose to follow the suggestions by Miller et al. (1993) and to conduct recalibration tests on the following null hypotheses: (i) $H_0$: $a = 0, b = 1$, which refers to an overall test for calibration; (ii) $H_0$: $a = 0 \mid b = 1$, to test for calibration-in-the-large given appropriate refinement and (iii) $H_0$: $b = 1 \mid a$, to test refinement given corrected calibration-in-the-large.

Because the predicted hazards $\hat{\lambda}_1$ are derived from a subdistribution hazard model that was fitted using weighted maximum likelihood estimation, we fit the recalibration model in (13) by optimizing a weighted binary log likelihood of the form

$$
\ell_{tc} = \sum_{m=1}^N \sum_{t=1}^{k-1} w_{mt} \{y_{mt} \log(\pi_1(t|x_m)) + (1 - y_{mt}) \log(1 - \pi_1(t|x_m))\},
$$

where the probabilities $\pi_1(t|x_m)$ are given by $\pi_1(t|x_m) = \exp(\eta_{tc}(t|x_m))/(1 + \exp(\eta_{tc}(t|x_m)))$. The binary outcome values $y_{mt}$ and the weights $w_{mt}$ are defined in the same way as in Section 4. Note that $\hat{V}(\cdot)$ is again estimated from the training sample with individuals $i = 1, \ldots, n$. In the case where $a = 0$ (referring to the tests in (i) and (ii) above), the log likelihood (14) can be written as

$$
\ell_{tc} = b \sum_{m=1}^N \sum_{t=1}^{k-1} w_{mt} y_{mt} \log (\hat{\lambda}_1(t|x_m))
$$
\[+ b \sum_{m=1}^{N} \sum_{t=1}^{k-1} w_{mt}(1 - y_{mt}) \log \left(1 - \hat{\lambda}_1(t|x_m)\right) \]
\[- N \sum_{m=1}^{k-1} w_{mt} \log \left(\hat{\lambda}_1(t|x_m) + (1 - \hat{\lambda}_1(t|x_m))^{b}\right)\]

which corresponds to the weighted log likelihood in equation (7) of Cox (1958). The derivation of (15) is given in Section 1 of the Supplementary Material. It follows that hypotheses (i)–(iii) can be examined using likelihood-ratio test statistics that asymptotically (as \(N \to \infty\)) follow \(\chi^2\)-distributions with one (hypotheses ii and iii) and two (hypothesis i) degrees of freedom, respectively.

6. NUMERICAL EXPERIMENTS

In this section we present the results of numerical experiments to evaluate the proposed calibration measures under known conditions. The main focus of the study was on measuring the performance in scenarios with different rates of type 1 events, different levels of censoring, a varying number of discrete time points, and various forms of misspecification.

6.1. Experimental Design

In order to generate data from a given subdistribution hazard model for type 1 events, we used a scheme adopted from Fine & Gray (1999). This procedure is also described in Beyersmann, Allignol & Schumacher (2011), where it was termed “indirect simulation.” In all simulation scenarios, we considered data with two competing events, \(e_i \in \{1, 2\}\), that was generated under the assumption of proportional subdistribution hazards. More specifically, our discrete subdistribution hazard model was based on the discretization of the continuous model

\[F_1(t|x_i) = P(T_{\text{cont},i} \leq t, e_i = 1|x_i) = 1 - (1 - q + q \exp(-t))^{\exp(x_i^\top r)},\]

where \(T_{\text{cont},i} \in \mathbb{R}^+\) denotes the continuous time span of individual \(i\), and \(r = (r_1, \ldots, r_p)^\top\) is a set of regression coefficients. The parameter \(q \in (0, 1)\) determines the probability of a type 1 event which, according to (16), was given by \(\pi_{i1} := P(e_i = 1|x_i) = 1 - (1 - q)^{\exp(x_i^\top r)}\). By definition, high values of \(q\) result in high probabilities of \(\pi_{i1}\), and vice versa. The probability of a competing event was given by \(\pi_{i2} := P(e_i = 2|x_i) = 1 - \pi_{i1} = (1 - q)^{\exp(x_i^\top r)}\).

Continuous time spans for type 2 events were drawn from the exponential model

\[T_{\text{cont},i}|e_i = 2, x_i \sim \text{Exp}(\xi_2 = \exp(x_i^\top \beta)),\]

where \(\beta = (\beta_1, \ldots, \beta_p)^\top\) denotes a set of regression coefficients linking the rate parameter \(\xi_2\) with the values of the covariates \(x\).

In order to obtain discrete event times \(T_{\text{disc},i}\), we generated data according to the indirect simulation scheme described above and grouped the resulting continuous event times into categories \(k \in \{5, 10, 15\}\). The latter were defined by the quantiles of the continuous event times, which were pre-estimated from an independent sample with 1,000,000 observations. As a consequence, the same interval boundaries were used in each simulation run. Censoring times were generated from a discrete distribution with probability density function \(P(C_{\text{disc},i} = t) = u^{k-t+1}/ \sum_{s=1}^{k} u^s, t = 1, \ldots, k\), where the percentage of censored observations was controlled by the parameter \(u \in \mathbb{R}^+\).
We considered two standard, normally distributed covariates \( x_{i1}, x_{i2} \sim \text{N}(0, 1) \) and two binary covariates \( x_{i3}, x_{i4} \sim \text{Bin}(1, 0.5) \). All covariates were independent, and the true regression coefficients were set to \( \gamma = (0.4, -0.4, 0.2, -0.2)^T \) and \( \beta = (-0.4, 0.4, -0.2, 0.2)^T \), see Fine & Gray (1999). We specified three different censoring rates, denoted by \textit{weak}, \textit{medium} and \textit{strong}, where the degree of censoring was controlled by the parameter \( u \) of the censoring distribution. More specifically, we used the values \( u = 0.85 \) (weak), \( u = 1 \) (medium) and \( u = 1.25 \) (strong), resulting in the censoring rates shown in Figure S1 of the Supplementary Material. We also considered three different probabilities of a type 1 event, specifying \( q \in \{0.2, 0.4, 0.8\} \). In total, this resulted in \( 3 \times 3 \times 3 = 27 \) different scenarios. All scenarios were analyzed using 100 replications with 5000 independently drawn observations each, which were equally split into a training sample and a validation sample \((n = 2500 \text{ and } N = 2500)\), respectively.

Figure S1 of the Supplementary Material illustrates the relative frequencies of observed events for the nine scenarios with \( k = 5 \). It is seen that the rates of observed type 1 events increased with increasing values of \( q \) and that censoring rates increased with increasing values of \( u \). For constant \( q \) and varying \( u \), the ratio of observed type 1 and type 2 events remained approximately the same. For \( q = 0.2 \) and \( q = 0.4 \), we observed more events of type 2 than of type 1, and for \( q = 0.8 \) there were more events of type 1 than type 2. For the scenarios with \( k = 10 \) and \( k = 15 \), the observed relative frequencies were almost the same and are thus not shown.

In all scenarios, the following models were fitted to the training samples: (a) the discrete subdistribution hazard model (5) with the inverse complementary log–log function (Gompertz model), which defines the same values of \( \gamma \) as the Fine and Gray proportional subdistribution hazards model in continuous time; (b) model (5) with a logistic distribution function \( h(\eta_i(t, x_i)) = \exp(\eta_i(t, x_i))/(1 + \exp(\eta_i(t, x_i))) \) and (c) a simple discrete hazard model (Gompertz model) for \( q = 0 \). In a fourth examination (d), we also fitted the discrete subdistribution hazard model as in (a), but slightly changed the data-generating process. The linear predictor \( x_i^T \beta \) in (16) was replaced by \( \sin(4x_{i1}) + \sin(4x_{i2}) + \gamma_3x_3 - \gamma_4x_4 \), which defines nonlinear effects of the two standard normally distributed covariates. Finally, we considered a setting (e), where the independence assumption between \( T_i \) and \( C_i \) was violated in the training sample. For this setting, we generated the censoring times from the discrete distribution given above with parameters

\[
u_i^{(e)} = \begin{cases} u + 0.25 & \text{if } T_{\text{disc},i} < \text{median}(T_{\text{disc}}), \\ u - 0.25 & \text{if } T_{\text{disc},i} \geq \text{median}(T_{\text{disc}}). \end{cases} \tag{17}
\]

According to the data-generating process defined by (e), the probability of censoring was much higher for observations with relatively small event times than for observations with medium or large event times.

In (a) we fitted the true data-generating model, whereas in (b)–(e) the fitted models were misspecified.

6.2. Results under Correct Model Specification

The calibration plots for one randomly chosen replication of the nine simulation scenarios with \( k = 5 \) when fitting the true data-generating model (a) are presented in Figure 1. Using Equation (12), the appropriate number of subsets in the scenario with \( q = 0.2 \) and weak censoring for this example was \( G = [20, 46] \). Thus, we set \( G = 20 \) in all calibration plots of the simulation study. The plots show that the empirical hazards \( \tilde{\lambda}_g \) and the average predicted hazards \( \frac{1}{G} \tilde{\lambda}_g \) coincided strongly, regardless of the degree of censoring and the rate of type 1 events. This result illustrates that the calibration plot defined in Section 4 is a strong tool for the graphical assessment of a correctly specified discrete subdistribution model. Figure 1 further shows that
modelling the subdistribution hazard $\lambda_1$ also works well in the presence of only a relatively small number of type 1 events (for $q = 0.2$ only about 10% type 1 events were observed). When fitting the logistic recalibration model, for example to the dataset with strong censoring and $q = 0.2$ (upper right panel of Figure 1), we obtained the estimates $\hat{a} = -0.084$ and $\hat{b} = 0.957$, which are close to the values $a = 0$ and $b = 1$ of perfect calibration. Exemplary calibration plots for the scenarios with $k = 10$ and $k = 15$ are presented in Figures S2 and S3 of the Supplementary Material. Again, the plots suggest nearly perfect calibration, with the exception of the scenarios with $k = 15$ and strong censoring, where the variation and thus the deviation from the 45-degree lines is more apparent.

The estimates of the calibration parameters $a$ and $b$ for all scenarios with $k = 5$ are shown in Figure 2. It is seen from the boxplots that, on average, the estimates were very close to values $a = 0$ and $b = 1$. In particular, for $q = 0.8$ (lower panel) the results of the recalibration model correctly indicated nearly perfect calibration. It is also seen that the variance of the estimates of the intercept $a$ increased with the decreasing rate of type 1 events. In contrast, the degree of censoring had only a small impact on the variance of the estimates. Figure 3 presents the corresponding $P$-values when conducting the recalibration tests (i)–(iii) specified in Section 5. Throughout all scenarios, the null hypotheses were kept in almost all replications (at the 5% type
Figure 2: Results of the simulation study when fitting the true data-generating model. The boxplots visualize the estimates of the calibration intercepts $a$ and calibration slopes $b$ that were obtained from fitting the logistic recalibration model ($k = 5$).

In particular, the tests for calibration-in-the-large given appropriate refinement (hypothesis ii) yielded very large $P$-values (corresponding to small negative log 10-transformed values). For example, in the scenario with $q = 0.2$ and strong censoring, this hypothesis was never rejected at the 5% type 1 error level. Overall, the results in Figures 2 and 3 illustrate that the proposed logistic recalibration model properly assessed the calibration of the fitted subdistribution hazard models, even in the case of strong censoring and a small rate of type 1 events.

The parameter estimates $\hat{a}$ and $\hat{b}$ and the $P$-values for the scenarios with $k = 10$ and $k = 15$ are given in Figures S4–S7 of the Supplementary Material. These results largely confirmed the previous findings for $k = 5$. Although the estimated calibration parameters deviated more strongly from $a = 0$ and $b = 1$, the associated null hypotheses were still kept at the 5% type 1 error level. The only exception was the scenario with $k = 15$, strong censoring, and $q = 0.2$ (upper right panel of Figure S7 of the Supplementary Material), where about half of the null hypotheses (i) and (iii) were rejected. These results, which were clearly related to the number of time intervals, can be explained by the fact that very few type 1 events were observed at later points in time when $k$ was increased. For example, with $k = 15$, strong censoring, and $q = 0.2$, fewer than four type 1 events occurred at time points $t > 10$ in most of the training and validation samples.
6.3. Results under Model Misspecification

In our second investigation (b), we fitted the discrete subdistribution hazard model with a logistic link function to the training samples. Regarding the calibration of the models, there was no noticeable difference to the correctly specified Gompertz model. Thus, the results were largely the same as those in Section 6.2 in all scenarios (not shown).

When fitting the simple discrete hazard model (c) to the training samples, the calibration of the models strongly deteriorated, which was clearly indicated by the proposed calibration measures. Exemplary calibration plots for the replication chosen in Figure 1 are shown in Figure 4. It is seen that, in particular, for the scenarios with weak and medium censoring, the set of points are mostly below the 45-degree line. Therefore, the predicted hazards were systematically too high. This result was also confirmed by the estimates of the recalibration intercepts \( \hat{a} \) (Figure S8 of the Supplementary Material), which were all below zero. In the scenarios with a small number of type 1 events \( (q = 0.2) \), the mean of estimates \( \hat{a} \) were smaller than \(-1\). Accordingly, the associated recalibration tests of the null hypotheses (i) and (ii) were consistently rejected (Figure S11 of the Supplementary Material). Rejection of the conditional null hypothesis (ii) again confirmed a systematic shift of the predicted hazards to higher values. Only in the scenario with strong censoring and \( q = 0.8 \) (lower right panel of Figure S11 of the Supplementary Material)
FIGURE 4: Results of the simulation study under the misspecified model (c). Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 10$). The 45-degree lines (dashed) indicate perfect calibration ($C = \text{degree of censoring}$).

Material), the recalibration tests still indicated good calibration. This result is clearly related to the small number of type 2 events, which did not substantially affect the fit of the simple discrete hazard model in this scenario.

Figure 5 depicts exemplary calibration plots obtained from the model fits under the third source of misspecification (d), where the predictor of the model was falsely specified to be linear. In comparison to Figure S2 of the Supplementary Material, the points are spread considerably wider around the 45-degree line. This result was confirmed by the estimated recalibration slopes $b$ (Figure S9 of the Supplementary Material), which were distinctly less than 1 (in particular, in the scenarios with $q = 0.2$). Remarkably, the test on null hypothesis (ii) was not affected by this form of misspecification throughout all scenarios (Figure S12 of the Supplementary Material). This demonstrates that calibration-in-the-large was still sufficient given appropriate refinement ($b = 1$). On the other hand, the two null hypotheses (i) and (iii) were more prone to the misspecification of the predictor function of the model, as they indicated poor calibration particularly in the scenarios with $q = 0.2$.

In the last setting (e) with violation of random censoring (calibration plots in Figure 6), the deviation from the 45-degree line is most evident in the scenarios with a high number of type 1 events ($q = 0.8$). This is also seen from the fits of the recalibration model (Figures S10 and S13 of the Supplementary Material), where the estimated coefficients were clearly too small and the three null hypotheses were rejected to a large extent. In contrast to the misspecification in (c),
FIGURE 5: Results of the simulation study under the misspecified model (d). Calibration plots refer to one randomly chosen replication in each simulation scenario using \( G = 20 \) subsets \((k = 10)\). The 45-degree lines (dashed) indicate perfect calibration (C = degree of censoring).

the scenario with strong censoring and \( q = 0.8 \) appeared to be most problematic. This is likely because the violation of random censoring mainly affects the estimation of \( \hat{V} \), which is even more inaccurate in the case of a small number of type 2 events.

To sum up, the findings in cases (c)–(e) demonstrate that the proposed calibration measures are sensitive to the severity of misspecification of the fitted models. Here, calibration issues were most pronounced for models with an incorrect form of the predictor function (and weak censoring), and when the random censoring assumption was violated (and the number of type 1 events was small). Note that, because it did not gain any further insight, we reduced our considerations to the scenarios with \( k = 10 \) in this section.

7. NP INFECTION IN INTENSIVE CARE UNITS

To illustrate the use of the proposed calibration measures, we validated the prediction model by Berger et al. (2020) by analyzing a dataset on the development of pneumonia, which is a common nosocomial, that is, hospital-acquired infection in intensive care units (ICUs). This dataset was also considered earlier by Beyersmann et al. (2006), Wolkewitz et al. (2008), and other authors. As NP has a strong impact on the mortality of patients in ICUs, it is of high interest to determine the risk factors for the development of the disease.

The data were collected for a prospective cohort study at five ICUs in one university hospital, lasting 18 months (from February 2000 to July 2001) and comprising \( n = 1876 \) patients with a
FIGURE 6: Results of the simulation study under the misspecified model (e). Calibration plots refer to one randomly chosen replication in each simulation scenario using $G = 20$ subsets ($k = 10$). The 45-degree lines (dashed) indicate perfect calibration ($C =$ degree of censoring).

duration of ICU stay of at least 2 days. The outcome of interest was the time to NP infection. Other possible events that were competing with the onset of NP (being the event of interest) were death and discharge from hospital alive. Owing to the study design, the observed event times were discrete, as they were measured on a daily basis. Berger et al. (2020) analyzed the data over a period of 60 days, resulting in 61 possible event times $t = 1, 2, \ldots, 61$, where $t = k = 61$ referred to all individuals with event times $\geq 61$ days. At the observed times, each patient acquired the NP infection ($n = 158$), died, was released from hospital ($n = 1695$), or was administratively censored ($n = 23$). Descriptive summary statistics of the baseline risk factors considered in the analysis were presented in Table 1 of Wolkewitz et al. (2008). In addition to the age of the patients (centred at 60 years), the gender of the patients, and the simplified acute physiology score (SAPS II), there were 11 binary risk factors characterizing the patients and their hospital stay. Note that SAPS II measures the severity of disease for patients admitted to ICUs aged 15 years or older. The score is calculated from 12 routine physiological measurements during the first 24 h, resulting in a range of [0, 163] points (Le Gall, Lemeshow & Saulnier, 1993). The binary variables either referred to the time of ICU admission (on admission) or the time prior to ICU admission (before admission).

We fitted the discrete subdistribution hazard model used in Berger et al. (2020) to the data described above. This model incorporates the baseline risk factors and a set of smooth baseline coefficients represented by cubic P-splines with a second-order difference penalty (fitted using
Analysis of the nosocomial pneumonia infection data. The calibration plot refers to a randomly chosen partition of the data into a training and a validation sample using \( G = 24 \) subsets. According to Equation (12), the appropriate number of subsets is \( G = \lfloor 24 \rfloor.21 \). The 45-degree line (dashed line) indicates perfect calibration.

This model was referred to as Model 2 in Berger et al. (2020). To assess the calibration of the model, we conducted a benchmark experiment which was based on 100 random partitions of the data. Each partition consisted of a training sample of size \( n = 1500 \) (80\%) and a validation sample of size \( N = 376 \) (20\%).

The main results in Berger et al. (2020) can be summarized as follows: risk factors significantly increasing the risk of NP acquisition at the 5\% type I error level were (i) male gender, (ii) an intubation on admission, (iii) no pneumonia on admission, (iv) another infection on admission, (v) an elective or emergency surgery before admission and (vi) a cardio/pulmonary or neurological underlying disease.

Figure 7 presents the calibration plot of the model that was obtained for one randomly chosen partition of the data. It is seen that apart from the three subsets defined by the largest percentiles, the empirical hazards \( \hat{y}_g \) and the average predicted hazards \( \bar{\lambda}_{1g} \) were very small (<0.005). Furthermore, the plot showed strong agreement between \( \hat{y}_g \) and \( \bar{\lambda}_{1g} \), indicating satisfactory calibration of the fitted model. The calibration plots obtained for 25 further partitions of the data are shown in Figure S14 of the Supplementary Material. Except single values, the plots do not reveal severe deviations from the 45-degree line. However, the bundle of small hazard values makes the evaluation of the plots rather difficult.

Boxplots of the estimated calibration parameters \( \hat{a} \) and \( \hat{b} \) and the \( P \)-values when performing the associated recalibration tests are shown in Figure 8. The estimates related to the calibration plot in Figure 7 were \( \hat{a} = 0.039 \) and \( \hat{b} = 0.984 \) with \( P \)-values 0.809 (hypothesis i), 0.518 (hypothesis ii), and 0.935 (hypothesis iii). The mean estimates of \( a \) and \( b \) (left panel of Figure 8) indicate that the predicted hazards tended to be too high \( (a < 0) \) and that they varied a little too much \( (0 < b < 1) \). Importantly, this trend was also seen in the simulations in Section 6.2 with weak censoring and \( q = 0.2 \) (see Figure 2 and Figures S4 and S5 of the Supplementary Material), which is the setting that is most comparable to the characteristics of the NP infection data. Also
FIGURE 8: Analysis of the NP infection data. Estimates of the calibration intercepts $a$ and calibration slopes $b$ (left) and negative log 10-transformed $P$-values obtained from the recalibration tests (right) obtained for 100 partitions of the data into training and validation samples.

Note that the number of observed type 1 events in the data is smaller than 3 at time points $t > 20$. According to the recalibration tests (right panel of Figure 8), the deviations of the calibration parameters from $a = 0$ and $b = 1$ were not substantial, as the majority of the null hypotheses on calibration-in-the-large and refinement were not rejected at the 5% type 1 error level. This result is again in line with the findings in the simulation study and also indicated a good calibration of the discrete subdistribution model derived in Berger et al. (2020).

8. CONCLUDING REMARKS

Discrete time-to-event models have gained widespread popularity in applied research in recent years (Tutz & Schmid, 2016; Lee, Feuer & Fine, 2018). Therefore, methodology for the proper validation of their generalization performance is increasingly necessary. In this regard, the methods presented here constitute a new set of tools to assess the calibration of discrete subdistribution hazard models for competing risks analysis. They consist of a calibration plot for graphical assessments as well as a recalibration model including tests on calibration-in-the-large and refinement. Both methods are well connected to analogous approaches for binary regression (Miller et al., 1993; Hosmer, Lemeshow & Sturdivant, 2013). In the single-event scenario, the graphical tool presented here naturally reduces to the calibration plot proposed in Berger & Schmid (2018).

Unlike Heyard et al. (2020), who proposed tools to assess the calibration of cause-specific hazard models, we considered the subdistribution framework originally proposed by Fine & Gray (1999) for competing risks data in continuous time. In contrast to cause-specific hazard modelling, this approach has the advantage that only one model needs to be considered if the interest is in the occurrence of one specific event. Subdistribution hazard modelling is of high practical importance, as it allows the interpretation of regression coefficients in terms of increasing/decreasing effects of the covariates on the incidence of the target event (Austin & Fine, 2017). Furthermore, Young et al. (2020) suggested using differences in the cumulative incidence functions as estimands in causal modelling. To evaluate the calibration of cumulative incidence functions, Lee (2017) generated an alternative kind of calibration plot that compared predictions of the cumulative incidence function to their respective nonparametric estimates.
The simulation study and the analysis of the NP infection data suggest that the methods work well under both correct and incorrect model specifications, even in “unfavourable” scenarios with a high censoring rate and few type 1 events. However, one should be careful in situations with a large number of time intervals when the observed number of type 1 events at later time points is rare.

All evaluations presented in this article were performed using the R add-on package discSurv (Welchowski & Schmid, 2019). It contains the function dataLongSubDist() to generate the binary outcome vectors (8) and the corresponding weights (9) and (10). Parameter estimates of the recalibration model were obtained by using the function glm() with the family function binomial() for logistic regression.

ACKNOWLEDGEMENTS

We thank Jan Beyersmann for fruitful discussions on subdistribution hazard modelling and for helpful suggestions on how to improve the manuscript. We thank the SIR-3 study investigators for providing us with the NP infection data. Support by the German Research Foundation (DFG), grant SCHM 2966/2-1, is gratefully acknowledged.

REFERENCES

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorioand, A., Devereaux, P. J., McGinn, T., & Guyatt, G. (2017). Discrimination and calibration of clinical prediction models: Users’ guides to the medical literature. *Journal of the American Medical Association*, 318, 1377–1384.

Andersen, P. K., Geskus, R. B., de Witte, T., & Putter, H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology*, 41, 861–870.

Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133, 601–609.

Austin, P. C. & Fine, J. P. (2017). Practical recommendations for reporting Fine–Gray model analyses for competing risk data. *Statistics in Medicine*, 36, 4391–4400.

Baker, S. G., Cook, N. R., Vickers, A., & Kramer, B. S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 729–748.

Berger, M. & Schmid, M. (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, 18, 322–345.

Beyersmann, J., Allignol, A., & Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. Springer, New York.

Braun, D., Gorfine, M., Katki, H. A., Ziogas, A., & Parmigiani, G. (2018). Nonparametric adjustment for measurement error in time-to-event data: Application to risk prediction models. *Journal of the American Statistical Association*, 113, 14–25.

Cortese, G. & Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, 52, 138–158.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.

Ding, A. A., Tian, S., Yu, Y., & Guo, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association*, 107, 990–1003.

Doane, D. P. (1976). Aesthetic frequency classifications. *The American Statistician*, 30, 181–183.

Fahrmeir, L. & Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association*, 91, 1584–1594.

Fine, J. P. & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94, 496–509.

DOI: 10.1002/cjs.11633
Gibbs, M. (2011). Ecological risk assessment, prediction, and assessing risk predictions. *Risk Analysis: An International Journal*, 31, 1784–1788.

Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer, New York.

Henderson, R. & Keiding, N. (2005). Individual survival time prediction using statistical models. *Journal of Medical Ethics*, 31, 703–706.

Heyard, R., Timsit, J.-F., Held, L., & COMBACTE-MAGNET Consortium. (2020). Validation of discrete time-to-event prediction models in the presence of competing risks. *Biometrical Journal*, 62, 643–657.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken, NJ, John Wiley & Sons.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, Hoboken.

Kerr, K. F. & Janes, H. (2017). First things first: Risk model performance metrics should reflect the clinical application. *Statistics in Medicine*, 36, 4503–4508.

Klein, J. P. & Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61, 223–229.

Lau, B., Cole, S. R., & Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American Journal of Epidemiology*, 170, 244–256.

Le Gall, J.-R., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270, 2957–2963.

Lee, M. (2017). Inference for cumulative incidence on discrete failure times with competing risks. *Journal of Statistical Computation and Simulation*, 87, 1989–2001.

Lee, M., Feuer, E. J., & Fine, J. P. (2018). On the analysis of discrete time competing risks data. *Biometrics*, 74, 1468–1481.

Liu, D., Zheng, Y., Prentice, R. L., & Hsu, L. (2014). Estimating risk with time-to-event data: An application to the women’s health initiative. *Journal of the American Statistical Association*, 109, 514–524.

Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., & McDonald, C. J. (1993). Validation of probabilistic predictions. *Medical Decision Making*, 13, 49–57.

Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M. (2012a). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, 98, 691–698.

Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012b). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98, 683–690.

Nightengale, F. (1863). *Notes on Hospitals*. Longman, Green, Longman, Roberts, and Green, London.

Poguntke, I., Schumacher, M., Beyersmann, J., & Wolkewitz, M. (2018). Simulation shows undesirable results for competing risks analysis with time-dependent covariates for clinical outcomes. *BMC Medical Research Methodology*, 18, 79.

Schmid, M. & Berger, M. (2020). Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics*, e1529. https://doi.org/10.1002/wics.1529.

Schmid, M., Tutz, G., & Welchezowski, T. (2018). Discrimination measures for discrete time-to-event predictions. *Econometrics and Statistics*, 7, 153–164.

Soave, D. M. & Strug, L. J. (2018). Testing calibration of Cox survival models at extremes of event risk. *Frontiers in Genetics*, 9, 177.

Steyerberg, E. W. (2019). *Clinical Prediction Models*, 2nd ed., Springer, New York.

Steyerberg, E. W. & Harrell, F. E. (2016). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, 69, 245–247.

Steyerberg, E. W. & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal*, 35, 1925–1931.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, 21, 128.

Tutz, G. & Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer, New York.

Vickers, A. J. & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26, 565–574.
Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352, i6.

Welchowski, T. & Schmid, M. (2019). *discSurv: Discrete Time Survival Analysis*. R package version 1.4.1. http://cran.r-project.org/web/packages/discSurv.

Witten, D. M. & Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19, 29–51.

Wolkewitz, M., Vonberg, R. P., Grundmann, H., Beyersmann, J., Gastmeier, P., Bärwolff, S., Geffers, C., Behnke, M., Rüden, H., & Schumacher, M. (2008). Risk factors for the development of nosocomial pneumonia and mortality on intensive care units: Application of competing risks models. *Critical Care*, 12, R44.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., & Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39, 1199–1236.

*Received 6 August 2020  
Accepted 22 December 2020*