The wild bootstrap for multivariate Nelson–Aalen estimators

Tobias Bluhmki¹ · Dennis Dobler² ·
Jan Beyersmann¹ · Markus Pauly¹

Received: 12 May 2017 / Accepted: 5 February 2018 / Published online: 6 March 2018
© The Author(s) 2018. This article is an open access publication

Abstract We rigorously extend the widely used wild bootstrap resampling technique to the multivariate Nelson–Aalen estimator under Aalen’s multiplicative intensity model. Aalen’s model covers general Markovian multistate models including competing risks subject to independent left-truncation and right-censoring. This leads to various statistical applications such as asymptotically valid confidence bands or tests for equivalence and proportional hazards. This is exemplified in a data analysis examining the impact of ventilation on the duration of intensive care unit stay. The finite sample properties of the new procedures are investigated in a simulation study.

Keywords Conditional central limit theorem · Counting process · Equivalence test · Proportional hazards · Kolmogorov–Smirnov test · Survival analysis · Weak convergence

1 Introduction

One of the most crucial quantities within the analysis of time-to-event data with independently right-censored and left-truncated survival times is the cumulative hazard

Tobias Bluhmki and Dennis Dobler are co-authors with equal contributions.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10985-018-9423-x>) contains supplementary material, which is available to authorized users.

✉ Dennis Dobler
d.dobler@vu.nl

¹ Institute of Statistics, Ulm University, Helmholtzstrasse 20, 89081 Ulm, Germany

² Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

function, also known as cumulative transition intensity. Most commonly, it is non-parametrically estimated by the well-known *Nelson–Aalen estimator* (Andersen et al. 1993, Chapter IV). In this context, time-simultaneous confidence bands are the perhaps best interpretative tool to account for related estimation uncertainties.

The construction of confidence bands is typically based on the asymptotic behavior of the underlying stochastic processes, more precisely, the (properly standardized) Nelson–Aalen estimator asymptotically behaves like a Wiener process. Early approaches utilized this property to derive confidence bands for the cumulative hazard function; see e.g., Bie et al. (1987) or Section IV.1.3 in Andersen et al. (1993).

However, Dudek et al. (2008) found that this approach applied to small samples can result in considerable deviations from the nominal level. To improve small sample properties, Efron (1979, 1981) suggested a computationally convenient and flexible resampling technique, called *bootstrap*, where the unknown non-Gaussian quantile is approximated via repeated generation of point estimates based on random samples of the original data. For a detailed discussion within the standard right-censored survival setup, see also Akritas (1986), Lo and Singh (1986), and Horvath and Yandell (1987). The simulation study of Dudek et al. (2008) particularly reports improvements of bootstrap-based confidence bands for the hazard function as compared to those using asymptotic quantiles. An alternative is the so-called *wild bootstrap* firstly proposed in the context of regression analyses (Wu 1986). As done in Lin et al. (1993), the basic idea is to replace the (standardized) residuals with independent standardized variates—so-called multipliers—while keeping the data fixed. One advantage compared to Efron’s bootstrap is to gain robustness against variance heteroscedasticity (Wu 1986). Using standard normal multipliers, this resampling procedure has been applied to construct time-simultaneous confidence bands for survival curves under the Cox proportional hazards model (Lin et al. 1994) and adapted to cumulative incidence functions in the competing risks setting (Lin 1997). The latter approach has recently been extended to general wild bootstrap multipliers with mean zero and variance one (Beyersmann et al. 2013), which indicate possible improved small sample performances. This result was confirmed in Dobler and Pauly (2014) as well as Dobler et al. (2017), where more general resampling schemes are discussed. In all these references, only one multiplier, typically standard normal, was required per individual, because each individual experienced at most one event. If an individual may experience more than one event, multiple multipliers per individual were, e.g., considered by Dabrowska and Ho (2000) in the context of Cox modelling, see also Shu et al. (2007) for Cox models in a semi-Markov illness-death model, and Liu et al. (2008) in a progressive multistate model for current leukemia free survival.

The present article focuses on the nonparametric estimation of cumulative hazard functions and proposes a general and flexible wild bootstrap resampling technique, which is valid for a large class of time-to-event models. In particular, the procedure is not limited to the standard survival or competing risks framework. The key assumption is that the involved counting processes satisfy the so-called *multiplicative intensity model* (Andersen et al. 1993). Consequently, arbitrary Markovian multistate models with finite state space are covered, as well as various other intensity models (e.g., excess or relative mortality models, cf. Andersen and Væth 1989) and specific semi-Markov

situations (Andersen et al. 1993, Example X.1.7). Independent right-censoring and left-truncation can straightforwardly be incorporated.

The main aim of this article is to mathematically justify the wild bootstrap technique for the multivariate Nelson–Aalen estimator in this general framework, while using not necessarily normally distributed but possibly multiple multipliers per counting process in the resampling step. This is accomplished by a novel martingale-based proof that discloses the close connection between the estimator and its wild bootstrap version. This insight would not have been possible by generalizing the elementary approach for showing bootstrap validity in competing risks given in Beyersmann et al. (2013) and Dobler and Pauly (2014) to the present set-up. Compared to the standard survival or competing risks setting, with at most one transition per individual, the major difficulty is to account for counting processes having an arbitrarily large random number of jumps. We will see that utilizing only one multiplier per individual counting process leads to the wrong covariance structure in general; instead, one multiplier per increment is required. As Beyersmann et al. (2013) suggested in the competing risks setting, we also allow for more general multipliers with expectation 0 and variance 1 and extend the resulting weak convergence theorems to resample the multivariate Nelson–Aalen estimator in our general setting. For practical applications, this result allows, for instance, within- or two-sample comparisons and the formulation of statistical tests.

The wild bootstrap is exemplified to statistically assess the impact of mechanical ventilation in the intensive care unit (ICU) on the length of stay. A related problem is to investigate ventilation-free days, which was established as an efficacy measure in patients subject to acute respiratory failure (Schoenfeld et al. 2002). However, applications of their methodology (see e.g., Sauaia et al. 2009; Stewart et al. 2009) rely on the constant hazards assumption. Other publications like de Wit et al. (2008), Trof et al. (2012), or Curley et al. (2015) used a Kaplan–Meier-type procedure that does not account for the more complex multistate structure. In contrast, we propose an illness-death model with recovery that methodologically works under the more general time-inhomogeneous Markov assumption and captures both the time-dependent structure of mechanical ventilation and the competing endpoint ‘death in ICU’.

The remainder of this article is organized as follows: Sect. 2 introduces cumulative hazard functions and their Nelson–Aalen estimators using counting process formulations. After summarizing its asymptotic properties, Sect. 3 offers our main theorem on conditional weak convergence for the wild bootstrap. This allows for various statistical applications in Sect. 4: Two-sided hypothesis tests and various sorts of time-simultaneous confidence bands are deduced, as well as simultaneous confidence intervals for a finite set of time points. Furthermore, tests for equivalence, inferiority and superiority as well as for proportionality of two hazard functions constitute useful criteria in practical data analyses. A simulation study assessing small and large sample performances of both the derived confidence bands in comparison to the algebraic approach based on the time-transformed Brownian motion and the tests for proportional hazards is reported in Sect. 5. The SIR-3 data on patients in ICU (Beyersmann et al. 2006; Wolkewitz et al. 2008) serves as its template and is practically revisited in Sect. 6. Concluding remarks and a discussion are given in Sect. 7. All proofs are deferred to Appendix A and the non-applicability of the ordinary multiplier resampling is verified in Appendix B.

2 Nonparametric estimation under the multiplicative intensity structure

Throughout, we adopt the notation of Andersen et al. (1993). For $k \in \mathbb{N}$, let $\mathbf{N} = (N_1, \dots, N_k)'$ be a multivariate counting process which is adapted to a filtration $(\mathcal{F}_t)_{t \geq 0}$. Each entry N_j , $j = 1, \dots, k$, is supposed to be a càdlàg function, zero at time zero, and to have piecewise constant paths with jumps of size one. In addition, assume that no two components jump at the same time and that each $N_j(t)$ satisfies the *multiplicative intensity model* of Aalen (1978) with intensity process given by $\lambda_j(t) = \alpha_j(t)Y_j(t)$. Here, $Y_j(t)$ defines a predictable process not depending on unknown parameters and α_j describes a non-negative (hazard) function. For well-definiteness, the observation of $\mathbf{N}$ is restricted to the interval $[0, \tau]$, where $\tau < \tau_j = \sup\{u \geq 0 : \int_{(0,u]} \alpha_j(s)ds < \infty\}$ for all $j = 1, \dots, k$. The multiplicative intensity structure covers several customary frameworks in the context of time-to-event analysis. The following overview specifies frequently used models.

Example 1 (a) Markovian multistate models with finite state space $\mathcal{S}$ are very popular in biostatistics. In this setting, $Y_\ell(t)$ represents the total number of individuals in state ℓ just prior to t ('number at risk'), whereas $\alpha_{\ell m}(t)$ is the instantaneous risk ('transition intensity') to switch from state ℓ to m , where $\ell, m \in \mathcal{S}$, $\ell \neq m$. Here, $N_\ell = \sum_{i=1}^n N_{\ell;i}$ is the aggregation over individual-specific counting processes with $n \in \mathbb{N}$ individuals under study. For specific examples (such as competing risks or the illness-death model) and details including the incorporation of independent left-truncation and right-censoring, see Andersen et al. (1993) and Aalen et al. (2008).

- (b) Other examples are the relative or excess mortality model, where not all individuals necessarily share the same hazard rate α . In this case Y cannot be interpreted as the total number of individuals at risk as in part (a); see Example IV.1.11 in Andersen et al. (1993) for details.
- (c) The time-inhomogeneous Markov assumption required in part (a) can even be relaxed in specific situations: Following Example X.1.7 in Andersen et al. (1993), consider an illness-death model without recovery. Assuming that the transition intensity α_{12} depends on the duration d in the intermediate state, but not on time t , leads to a semi-Markov process not satisfying the multiplicative intensity structure. This is because the intensity process of $N_{12}(t)$ is given by $\alpha_{12}(t - T)Y_1(t)$, where the first factor of the product is not deterministic anymore. Here, T is the random transition time into state 1. However, when $d = t - T$ is used as the basic timescale, the counting process $K(d) = N_{12}(d + T)$ has intensity $\alpha_{12}(d)Y_1(d)$ with respect to the filtration

$$\mathcal{F}_d = (\sigma\{(N_{01}(t), N_{02}(t)) : 0 < t < \tau\} \vee \sigma\{K(d) : 0 < d < \infty\}).$$

Thus, the multiplicative intensity structure is fulfilled.

Under the above assumptions, the Doob–Meyer decomposition applied to N_j leads to

$$dN_j(s) = \lambda_j(s)ds + dM_j(s), \quad (2.1)$$

where the M_j are zero-mean martingales with respect to $(\mathcal{F}_t)_{t \in [0, \tau]}$. The canonical nonparametric estimator of the cumulative hazard function $A_j(t) = \int_{(0, t]} \alpha_j(s) ds$ is given by the so-called *Nelson–Aalen estimator*

$$\hat{A}_{jn}(t) = \int_{(0, t]} \frac{J_j(s)}{Y_j(s)} dN_j(s).$$

Here, $J_j(t) = \mathbf{1}\{Y_j(t) > 0\}$, $\frac{0}{0} := 0$, and $n \in \mathbb{N}$ is a sample size-related number (that goes to infinity in asymptotic considerations). Its multivariate counterpart is introduced by $\hat{\mathbf{A}}_n := (\hat{A}_{1n}, \dots, \hat{A}_{kn})'$. As in Andersen et al. (1993), suppose that there exist deterministic functions y_j with $\inf_{u \in [0, \tau]} y_j(u) > 0$ such that

$$\sup_{s \in [0, \tau]} \left| \frac{Y_j(s)}{n} - y_j(s) \right| \xrightarrow{P} 0 \quad \text{for all } j = 1, \dots, k, \quad (2.2)$$

where ‘ $\xrightarrow{P}$ ’ denotes convergence in probability for $n \rightarrow \infty$. For each j , define the normalized Nelson–Aalen process $W_{jn} := \sqrt{n}(\hat{A}_{jn} - A_j)$ possessing the asymptotic martingale representation

$$W_{jn}(t) \doteq \sqrt{n} \int_{(0, t]} \frac{J_j(s)}{Y_j(s)} dM_j(s) \quad (2.3)$$

with M_j given by (2.1). Here, ‘ $\doteq$ ’ means that the difference of both sides converges to zero in probability. Define the vectorial aggregation of all W_{jn} as $\mathbf{W}_n = (W_{1n}, \dots, W_{kn})'$ and let ‘ $\xrightarrow{d}$ ’ denote convergence in distribution for $n \rightarrow \infty$. Then, Theorem IV.1.2 in Andersen et al. (1993) in combination with (2.2) provides a weak convergence result on the k -dimensional space $\mathcal{D}[0, \tau]^k$ of càdlàg functions endowed with the product Skorohod topology.

Theorem 1 *If assumption (2.2) holds, we have convergence in distribution*

$$\mathbf{W}_n \xrightarrow{d} \mathbf{U} = (U_1, \dots, U_k)', \quad (2.4)$$

on $\mathcal{D}[0, \tau]^k$, where $U_1, \dots, U_k$ are independent zero-mean Gaussian martingales with covariance functions $\psi_j(s_1, s_2) := \text{Cov}(U_j(s_1), U_j(s_2)) = \int_{(0, s_1]} \frac{\alpha_j(s)}{y_j(s)} ds$ for $j = 1, \dots, k$ and $0 \leq s_1 \leq s_2 \leq \tau$.

The covariance function ψ_j is commonly approximated by the *Aalen-type*

$$\hat{\sigma}_j^2(s_1) = n \int_{(0, s_1]} \frac{J_j(s)}{Y_j^2(s)} dN_j(s). \quad (2.5)$$

or the *Greenwood-type* estimator

$$\hat{\sigma}_j^2(s_1) = n \int_{(0, s_1]} \frac{J_j(s)(Y_j(s) - \Delta N_j(s))}{Y_j^3(s)} dN_j(s) \quad (2.6)$$

which are consistent for $\psi_j(s_1, s_2)$ under the assumption of Theorem 1; cf. (4.1.6) and (4.1.7) in Andersen et al. (1993). Here, $\Delta N_j(s)$ denotes the jump size of N_j at time s .

3 Inference via Brownian bridges and the wild bootstrap

As discussed in Andersen et al. (1993), the limit process U can analytically be approximated via Brownian bridges. However, improved coverage probabilities in the simulation study in Sect. 5 suggest that the proposed wild bootstrap approach may be preferable. First, we sum up the classic result.

3.1 Inference via transformed Brownian bridges

The asymptotic mutual independence stated in Theorem 1 allows to focus on a single component of W_n , say $W_{1n} = \sqrt{n}(\hat{A}_{1n} - A_1)$. For notational convenience, we suppress the subscript 1. Let g be a positive (weight) function on an interval $[t_1, t_2] \subset [0, \tau]$ of interest and B^0 a standard Brownian bridge process. Then, as $n \rightarrow \infty$, it is established in Section IV.1 in Andersen et al. (1993) that

$$\sup_{s \in [t_1, t_2]} \left| \frac{\sqrt{n}(\hat{A}_n(s) - A(s))}{1 + \hat{\sigma}^2(s)} g\left(\frac{\hat{\sigma}^2(s)}{1 + \hat{\sigma}^2(s)}\right) \right| \xrightarrow{d} \sup_{s \in [\phi(t_1), \phi(t_2)]} |g(s)B^0(s)|. \quad (3.1)$$

Here $\phi(t) = \frac{\sigma^2(t)}{1 + \sigma^2(t)}$, $\sigma^2(t) = \psi(t, t)$ and $\hat{\sigma}^2(t)$ is a consistent estimator for $\sigma^2(t)$, such as (2.5) or (2.6). Quantiles of the right-hand side of (3.1) for $g \equiv 1$ are recorded in tables (e.g., Koziol and Byar 1975; Hall and Wellner 1980; Schumacher 1984). For general g , they can be approximated via standard statistical software.

Even though relation (3.1) enables statistical inference based on the asymptotics of a central limit theorem, appropriate resampling procedures usually showed improved properties; see e.g., Hall and Wilson (1991), Good (2005) and Pauly et al. (2015).

3.2 Wild bootstrap resampling

In contrast to, for instance, a competing risks model where each counting process N_j is at most n , the number $N_j(\tau)$ is not necessarily bounded in our setup only assuming Aalen's multiplicative intensity model. Hence, a modification of the multiplier resampling scheme under competing risks suggested by Lin (1997) and elaborated by Beyersmann et al. (2013) is required. For this purpose, introduce counting process-specific stochastic processes indexed by $s \in [0, \tau]$ that are independent of N_j, Y_j for all $j = 1, \dots, k$. Let $(G_j(s))_{s \in [0, \tau]}$, $1 \leq j \leq k$, be independently and identically

distributed (i.i.d.) white noise processes such that each $G_j(s)$ satisfies $\mathbb{E}(G_j(s)) = 0$ and $\text{var}(G_j(s)) = 1$, $j = 1, \dots, k$, $s \in [0, \tau]$. That is, all ℓ -dimensional marginals of G_1 , $\ell \in \mathbb{N}$, shall be the same ℓ -fold product-measure. Then, a *wild bootstrap version* of the normalized multivariate Nelson–Aalen estimator $\mathbf{W}_n$ is defined as

$$\begin{aligned}\hat{\mathbf{W}}_n(t) &= (\hat{W}_{1n}(t), \dots, \hat{W}_{kn}(t))' \\ &:= \sqrt{n} \left(\int_{(0,t]} \frac{J_1(s)}{Y_1(s)} G_1(s) dN_1(s), \dots, \int_{(0,t]} \frac{J_k(s)}{Y_k(s)} G_k(s) dN_k(s) \right)'. \quad (3.2)\end{aligned}$$

In words, $\hat{\mathbf{W}}_n$ is obtained from representation (2.3) of $\mathbf{W}_n$ by substituting the unknown individual martingale processes M_j with the *observable* quantities $G_j N_j$. Even though only the values of each G_j at the jump times of N_j are relevant, this construction in terms of white noise processes enables a consideration of the wild bootstrap process on a product probability space; see the Appendix for details.

Consider for a moment the special case of a multistate model with n i.i.d. individuals (Example 1(a)). For instance, the competing risks model in Lin (1997) involves at most one transition (and thus one multiplier) per individual, while Glidden (2002) introduces only one multiplier per individual for estimating state occupation probabilities in general non-Markov multistate models. In contrast, our resampling approach is a new approach in the sense that it involves independent weightings of all jumps even within the same individual. The consequence is that, instead of considering one multiplier per individual, we need to utilize a white noise processes as done in (3.2) in order to account for randomly many numbers of events per individual.

The limit distribution of $\hat{\mathbf{W}}_n$ may be approximated by simulating a large number of replicates of the G 's, while the data is kept fixed. For a competing risks setting with standard normally distributed multipliers, our general scheme reduces to the one discussed in Lin (1997).

For the remainder of the paper, we summarize the available data in the σ -algebra $\mathcal{C}_0 = \sigma\{N_j(u), Y_j(u) : j = 1, \dots, k, u \in [0, \tau]\}$. A natural way to introduce a filtration based on $\mathcal{C}_0$ that progressively collects information on the white noise processes is by setting

$$\mathcal{C}_t = \mathcal{C}_0 \vee \sigma\{G_j(s) : j = 1, \dots, k, s \in [0, t]\}.$$

The following lemma is a key argument in an innovative, martingale-based consistency proof of the proposed wild bootstrap technique.

Lemma 1 *For each $n \in \mathbb{N}$, the wild bootstrap version of the multivariate Nelson–Aalen estimator $(\hat{\mathbf{W}}_n(t))_{t \in [0, \tau]}$ is a square-integrable martingale with respect to the filtration $(\mathcal{C}_t)_{t \in [0, \tau]}$ with orthogonal components. Its predictable variation process is given by*

$$\langle \hat{\mathbf{W}}_n \rangle : t \mapsto n \left(\int_0^t \frac{J_1(s)}{Y_1^2(s)} dN_1(s), \dots, \int_0^t \frac{J_k(s)}{Y_k^2(s)} dN_k(s) \right)$$

and its optional variation process by

$$[\hat{\mathbf{W}}_n] : t \mapsto n \left(\int_0^t \frac{J_1(s)}{Y_1^2(s)} G_1^2(s) dN_1(s), \dots, \int_0^t \frac{J_k(s)}{Y_k^2(s)} G_k^2(s) dN_k(s) \right).$$

The following conditional weak convergence result justifies the approximation of the limit distribution of $\mathbf{W}_n$ via $\hat{\mathbf{W}}_n$ given $\mathcal{C}_0$. Both, the general framework requiring only Aalen's multiplicative intensity structure as well as using possibly non-normal multipliers are original to the present paper.

Theorem 2 *Let U be as in Theorem 1. Assuming (2.2), we have the following conditional convergence in distribution on $\mathfrak{D}[0, \tau]^k$ given $\mathcal{C}_0$ as $n \rightarrow \infty$:*

$$\hat{\mathbf{W}}_n \xrightarrow{d} U \text{ in probability.}$$

Remark 1 (a) Reconsider the ordinary multiplier resampling based on a sequence of (time-constant) i.i.d. random variables $D_1, \dots, D_n$ with $E(D_1) = 0$ and unit variance where, in the resampling step, the martingales M_j are replaced with $D_j N_j$. In contrast to the wild bootstrap based on white noise processes, the wild bootstrap using the time-constant sequence $D_1, \dots, D_n$ fails to reproduce the correct covariance structure of the Nelson–Aalen process. Even in the special univariate Markovian case, the limit process does not have independent increments and it hence necessarily differs from the asymptotics described in Theorem 2; see Appendix B for details.

- (b) It is due to the martingale property of the wild bootstrapped multivariate Nelson–Aalen estimator that we anticipate a good finite sample approximation of the unknown distribution of the Nelson–Aalen estimator. In particular, the wild bootstrap, realized by white noise processes as above, succeeds in imitating the martingale structure of the original Nelson–Aalen estimator. The predictable variation process of the wild bootstrap process equals the optional variation process of the centered Nelson–Aalen process. Hence, both processes share the same properties and approximately the same covariance structure.
- (c) Suppose that $E(n^k J_1(u)/Y_1^k(u)) = O(1)$ for some $k \in \mathbb{N}$ and all $u \in [0, \tau]$, which for example holds for any $k \in \mathbb{N}$ if Y_1 has a number at risk interpretation. Since different increments of $\mathbf{W}_n$ (to arbitrary powers) are uncorrelated, it can be shown that the convergence in Theorem 1 for single $t \in [0, \tau]$ even holds in the Mallows metric d_p for any $0 < p \leq k$; see, for instance, Bickel and Freedman (1981) for such theorems related to the classical bootstrap. Provided that the r th moment of $G_1(u)$ exists, similar arguments show that the convergence in probability in Theorem 2 for single $t \in [0, \tau]$ holds in the Mallows metric d_p for any even $0 < p \leq r$ as well. This of course includes white noise processes with centered $Poi(1)$ or standard normal marginals, as applied later on.

4 Statistical applications

Throughout this section denote by $\alpha \in (0, 1)$ the nominal level of all inference procedures.

4.1 Confidence bands

After having established all required weak convergence results, we discuss different possibilities for realizing confidence bands for A_j around the Nelson–Aalen estimator $\hat{A}_{jn}$, $j = 1, \dots, k$, on an interval $[t_1, t_2] \subset [0, \tau]$ of interest. Later on, we propose a confidence band for differences of cumulative hazard functions. As in Sect. 3.1, we first focus on A_1 and suppress the index 1 for notational convenience. Following Andersen et al. (1993), Section IV.1, we consider weight functions

$$g_1(s) = (s(1-s))^{-1/2} \quad \text{or} \quad g_2 \equiv 1$$

as choices for g in relation (3.1). The resulting confidence bands are commonly known as *equal precision* and *Hall–Wellner* bands, respectively. We apply a log-transformation in order to improve small sample level α control. Combining the previous sections’ convergences with the functional delta-method and Slutsky’s lemma yields

Theorem 3 *Under condition (2.2), for any $0 \leq t_1 \leq t_2 \leq \tau$ such that $A(t_1) > 0$, we have the following convergences in distribution on the càdlàg space $\mathfrak{D}[t_1, t_2]$:*

$$\left(\sqrt{n} \hat{A}_n \frac{\log \hat{A}_n - \log A}{1 + \hat{\sigma}^2} \right) \cdot g \circ \frac{\hat{\sigma}^2}{1 + \hat{\sigma}^2} \xrightarrow{d} (gB^0) \circ \phi \quad \text{and} \quad (4.1)$$

$$\left(\frac{\hat{W}_n}{1 + \sigma^{*2}} \right) \cdot g \circ \frac{\sigma^{*2}}{1 + \sigma^{*2}} \xrightarrow{d} (gB^0) \circ \phi \quad (4.2)$$

conditionally given $\mathcal{C}_0$ in probability, with ϕ as in Sect. 3 and the wild bootstrap variance estimator $\sigma^{*2}(s) := n \int_{(0,t]} J(s) Y^{-2}(s) G^2(s) dN(s)$.

In particular, σ^{*2} is a uniformly consistent estimate for σ^2 (Dobler and Pauly 2014) and, being the optional variation process of the wild bootstrap Nelson–Aalen process, it may be one natural choice for variance estimation. For practical purposes, we adapt the approach of Beyersmann et al. (2013) and estimate σ^2 based on the empirical variance of the wild bootstrap quantities $\hat{W}_n$. The continuity of the supremum functional translates (4.1) and (4.2) into weak convergences for the corresponding suprema. Hence, the consistency of the following critical values is ensured:

$$c_{1-\alpha}^g = (1 - \alpha) \text{ quantile of } \mathfrak{L} \left(\sup_{s \in [t_1, t_2]} |g(\hat{\phi}(s)) B^0(\hat{\phi}(s))| \right),$$

$$\tilde{c}_{1-\alpha}^g = (1 - \alpha) \text{ quantile of } \mathfrak{L} \left(\sup_{s \in [t_1, t_2]} \left\| \frac{\hat{W}_n(s)}{1 + \sigma^{*2}(s)} g \left(\frac{\sigma^{*2}(s)}{1 + \sigma^{*2}(s)} \right) \right\| \mathcal{C}_0 \right),$$

where $\mathfrak{L}(\cdot)$ denotes the law of a random variable. Here, g equals either g_1 or g_2 and $\hat{\phi} = \frac{\hat{\sigma}^2}{1+\hat{\sigma}^2}$. Note, that $\tilde{c}_{1-\alpha}^g$ is, in fact, a random variable. The results are back-transformed into four confidence bands for A abbreviated with HW and EP for the Hall–Wellner and equal precision bands and a and w for bands based on quantiles of the asymptotic distribution and the wild bootstrap, respectively. In our simulation studies these bands are also compared with the linear confidence band CB_{dir}^w , which is based on the critical value

$$\tilde{c}_{1-\alpha} = (1 - \alpha) \text{ quantile of } \mathfrak{L}\left(\sup_{s \in [t_1, t_2]} |\hat{W}_n(s)| \middle| C_0\right).$$

Corollary 1 *Under the assumptions of Theorem 3, the following bands for the cumulative hazard function $(A(s))_{s \in [t_1, t_2]}$ provide an asymptotic coverage probability of $1 - \alpha$:*

$$\begin{aligned} CB_{EP}^a &= \left[\hat{A}_n(s) \exp \left(\mp \frac{c_{1-\alpha}^{g_1}}{\sqrt{n} \hat{A}_n(s)} \hat{\sigma}_n(s) \right) \right]_{s \in [t_1, t_2]} \\ CB_{HW}^a &= \left[\hat{A}_n(s) \exp \left(\mp \frac{c_{1-\alpha}^{g_2}}{\sqrt{n} \hat{A}_n(s)} (1 + \hat{\sigma}_n^2(s)) \right) \right]_{s \in [t_1, t_2]} \\ CB_{EP}^w &= \left[\hat{A}_n(s) \exp \left(\mp \frac{\tilde{c}_{1-\alpha}^{g_1}}{\sqrt{n} \hat{A}_n(s)} \hat{\sigma}_n(s) \right) \right]_{s \in [t_1, t_2]} \\ CB_{HW}^w &= \left[\hat{A}_n(s) \exp \left(\mp \frac{\tilde{c}_{1-\alpha}^{g_2}}{\sqrt{n} \hat{A}_n(s)} (1 + \hat{\sigma}_n^2(s)) \right) \right]_{s \in [t_1, t_2]} \\ CB_{dir}^w &= \left[\hat{A}_n(s) \mp \frac{\tilde{c}_{1-\alpha}}{\sqrt{n}} \right]_{s \in [t_1, t_2]}. \end{aligned} \quad (4.3)$$

Remark 2 1. Note that the wild bootstrap quantile $\tilde{c}_{1-\alpha}$ does not require an estimate of ϕ , thereby eliminating one possible cause of inaccuracy within the derivation of the other bands. However, the corresponding band CB_{dir}^w has the disadvantage to possibly include negative values.

2. The confidence bands are only well-defined if the left endpoint t_1 of the bands' time interval is larger than the first observed event. In particular, these bands yield unstable results for small values of $\hat{A}_n(t_1)$ due to the division in the exponential function; see Lin et al. (1994) for a similar observation.
3. The present approach directly allows the construction of confidence bands for within-sample comparisons of multiple $A_1, \dots, A_k$. For instance, a confidence band for the difference $A_1 - A_2$ may be obtained via quantiles based on the conditional convergence in distribution $\hat{W}_{1n} - \hat{W}_{2n} \xrightarrow{d} U_1 - U_2 \sim \text{Gauss}(0, \psi_1 + \psi_2)$ in probability by simply applying the continuous mapping theorem and taking advantage of the independence of U_1 and U_2 ; see Whitt (1980) for the continuity of the difference functional. For that purpose, the distribution of

$$D(t) = \sqrt{n}g(t)(\hat{A}_{1n}(t) - A_1(t) - (\hat{A}_{2n}(t) - A_2(t))), \quad (4.4)$$

with positive weight function g can be approximated by the conditional distribution of $\hat{D}(t) = g(t)(\hat{W}_{1n}(t) - \hat{W}_{2n}(t))$. With $g \equiv 1$, an approximate $(1 - \alpha) \cdot 100\%$ confidence band for the difference $A_1 - A_2$ of two cumulative hazard functions on $[t_1, t_2]$ is

$$\left[(\hat{A}_1(s) - \hat{A}_2(s)) \pm \tilde{q}_{1-\alpha}/\sqrt{n} \right]_{s \in [t_1, t_2]}, \quad (4.5)$$

where

$$\tilde{q}_{1-\alpha} = (1 - \alpha) \text{ quantile of } \mathfrak{L}\left(\sup_{s \in [t_1, t_2]} |\hat{W}_{1n}(s) - \hat{W}_{2n}(s)| \mid \mathcal{C}_0\right).$$

Similar arguments additionally enable common two-sample comparisons. A practical data analysis using other weight functions g in the context of cumulative incidence functions is given in Hieke et al. (2013).

Remark 3 (Construction of confidence intervals)

1. In particular, Theorem 3 yields a convergence result on $\mathbb{R}^m$ for a finite set of time points $\{s_1, \dots, s_m\} \subset [0, \tau]$, $m \in \mathbb{N}$. Hence, using critical values $\tilde{c}_{1-\alpha}$ and $\tilde{c}_{1-\alpha}^g$ obtained from the law of the maximum $\max_{s_1, \dots, s_m}$ instead of the supremum, a variant of Corollary 1 specifies simultaneous confidence intervals $I_1 \times \dots \times I_m$ for $(A(s_1), \dots, A(s_m))$ with asymptotic coverage probability $1 - \alpha$. Since the error multiplicity is taken into account, the asymptotic coverage probability of a single such interval I_j for $A(s_j)$ is greater than $1 - \alpha$.
2. Due to the asymptotic independence of the entries of the multivariate Nelson–Aalen estimator, a confidence region for the value of a multivariate cumulative hazard function $(A_1(t), \dots, A_k(t))$ at time $t \in [0, \tau]$ may be found using Šidák’s correction: Letting $J_1, \dots, J_k$ be pointwise confidence intervals for $A_1(t), \dots, A_k(t)$ with asymptotic coverage probability $(1 - \alpha)^{1/k}$, each found using the wild bootstrap principle, the coverage probability of $J_1 \times \dots \times J_k$ for $A_1(t) \times \dots \times A_k(t)$ clearly goes to $1 - \alpha$ as $n \rightarrow \infty$.

4.2 Hypothesis tests for equivalence, inferiority, superiority, and equality

Adapting the principle of confidence interval inclusion as discussed in Wellek (2010), Section 3.1, to time-simultaneous confidence bands, hypothesis tests for equivalence of cumulative hazard functions become readily available. To this end, let $\ell, u : [t_1, t_2] \rightarrow (0, \infty)$ be positive, continuous functions and denote by $(a_n(s), \infty)_{s \in [t_1, t_2]}$ and $[0, b_n(s))_{s \in [t_1, t_2]}$ the one-sided (half-open) analogues of any confidence band of the previous subsection with asymptotic coverage probability $1 - \alpha$. Furthermore, let $A_0 : [t_1, t_2] \rightarrow [0, \infty)$ be a pre-specified non-decreasing, continuous function for which equivalence to A shall be tested. More precisely:

$$H : \{A(s) \leq A_0(s) - \ell(s) \text{ or } A(s) \geq A_0(s) + u(s) \text{ for some } s \in [t_1, t_2]\} \\ \text{vs. } K : \{A_0(s) - \ell(s) < A(s) < A_0(s) + u(s) \text{ for all } s \in [t_1, t_2]\}.$$

Corollary 2 *Under the assumptions of Theorem 3, a hypothesis test ψ_n of asymptotic level α for H vs K is given by the following decision rule: Reject H if and only if the combined two-sided confidence band $(a_n(s), b_n(s))_{s \in [t_1, t_2]}$ is fully contained in the region spanned by $(A_0(s) - \ell(s), A_0(s) + u(s))_{s \in [t_1, t_2]}$. Further, it holds under K that $\mathbb{E}(\psi_n) \rightarrow 1$ as $n \rightarrow \infty$, i.e., ψ_n is consistent.*

Similar arguments lead to analogue one-sided tests for the inferiority or superiority of the true cumulative hazard function to a prespecified function A_0 . Moreover, statistical tests for equality of two cumulative hazard functions can be constructed using the weak convergence results of Remark 2(c):

$$H_ = \{A_1 \equiv A_2 \text{ on } [t_1, t_2]\} \quad \text{vs} \quad K_ = \{A_1(s) \neq A_2(s) \text{ for some } s \in [t_1, t_2]\}.$$

Corollary 3 below yields an asymptotic level α test for $H_ =$. Bajorunaite and Klein (2007) and Dobler and Pauly (2014) used similar two-sided tests for comparing cumulative incidence functions in a two-sample problem.

Corollary 3 (A Kolmogorov–Smirnov-type test) *Under the assumptions of Theorem 3 and letting g again be a positive weight function,*

$$\varphi_n^{KS} = \mathbf{1}\left\{ \sup_{s \in [t_1, t_2]} \sqrt{n}g(s) |\hat{A}_{1n}(s) - \hat{A}_{2n}(s)| > \tilde{q}_{1-\alpha} \right\}$$

defines a consistent, asymptotic level α resampling test for $H_ =$ vs. $K_ =$. Here $\tilde{q}_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of $\mathcal{L}\left(\sup_{s \in [t_1, t_2]} |\hat{D}(s)| \mid \mathcal{C}_0\right)$.

Similarly, Theorem 3 enables the construction of other tests, e.g., such of Cramér-von Mises-type. Furthermore, by taking the suprema over a discrete set $\{s_1, \dots, s_m\} \subset [0, \tau]$, the Kolmogorov–Smirnov test of Corollary 3 can also be used to test

$$\tilde{H}_ = \{A_1(s_j) = A_2(s_j) \text{ for all } 1 \leq j \leq m\} \\ \text{vs. } \tilde{K}_ = \{A_1(s_j) \neq A_2(s_j) \text{ for some } 1 \leq j \leq m\}.$$

Note that in a similar way, two-sample extensions of Corollaries 2 and 3 can be established following Dobler and Pauly (2014).

4.3 Tests for proportionality

A major assumption of the widely used Cox (1972) regression model is the assumption of proportional hazards over time. Several authors have developed procedures for testing the null hypothesis of proportionality, see e.g., Gill and Schumacher (1987), Lin (1991), Grambsch and Therneau (1994), Hess (1995), Scheike and Martinussen (2004), Kraus (2007), Bagdonavičius et al. (2010), or Chen et al. (2015), and the

references cited therein. However, most of these approaches have mainly been investigated for the standard survival framework under (independent) right-censoring and left-truncation mechanisms, even though they may be generalized to more general settings. In the context of the proposed wild bootstrap resampling technique, this motivates the explicit formulation of a two-sample proportionality test for the general setting only assuming a multiplicative intensity structure. This covers, for instance, arbitrary Markovian multistate models.

The framework is an unpaired two-sample model given by independent counting processes $N^{(1)}, N^{(2)}$ and predictable processes $Y^{(1)}, Y^{(2)}$, assuming the conditions of Sect. 2 for each group, and with sample sizes n_1 and n_2 , respectively. Let again $J^{(j)}(t) = \mathbf{1}\{Y^{(j)}(t) > 0\}$, $j = 1, 2$. Denote by $\hat{A}_{n_j}^{(j)} = \int_{(0,t]} \frac{J^{(j)}(s)}{Y^{(j)}(s)} dN^{(j)}$ the Nelson–Aalen estimator of the cumulative hazard functions $A^{(j)}$ and by $\alpha^{(j)}$ the corresponding rates, $j = 1, 2$. To motivate a suitable test statistic we make use of the following equivalence between hazards proportionality and equality of both cumulative hazards:

$$\begin{aligned} \alpha^{(1)}(t) &= c \alpha^{(2)}(t) \text{ in } t \in [0, \tau] \text{ for } c > 0 \\ \iff A^{(1)}(t) &= c A^{(2)}(t) \text{ in } t \in [0, \tau] \text{ for } c > 0, \end{aligned}$$

which, as the null hypothesis of interest, is denoted by $H_{0,\text{prop}}$. In a natural way, similar to Gill and Schumacher (1987), this leads to statistics of the form

$$T_{n_1, n_2} = \rho \left(\sqrt{\frac{n_1 n_2}{n}} \frac{\hat{A}_{n_2}^{(2)}}{\hat{A}_{n_1}^{(1)}}, \sqrt{\frac{n_1 n_2}{n}} \frac{\hat{A}_{n_2}^{(2)}(\tau)}{\hat{A}_{n_1}^{(1)}(\tau)} \right),$$

$n = n_1 + n_2$, where ρ is an adequate distance on $\mathfrak{D}[0, \tau]$, e.g., $\rho(f, g) = \sup w |f - g|$ (leading to Kolmogorov–Smirnov-type tests), $\rho(f, g) = \int (f - g)^2 w^2 d\lambda$ (leading to Cramér–von-Mises-type tests), where $w : [0, \tau] \rightarrow [0, \infty)$ is a suitable weight function. Later on, we choose $w = \hat{A}_{n_1}^{(1)}$ which ensures the evaluation of ρ on $\{\hat{A}_{n_1}^{(1)} > 0\}$. Let $\hat{W}_{n_1}^{(1)}$ and $\hat{W}_{n_2}^{(2)}$ be the obvious wild bootstrap versions of the sample-specific centered Nelson–Aalen estimators; cf. (3.2).

Theorem 4 *Let ρ be either the above Kolmogorov–Smirnov- or the Cramér–von Mises-type statistic with $w = \hat{A}_{n_1}^{(1)}$. If $n_1/n \rightarrow p \in (0, 1)$ as $\min(n_1, n_2) \rightarrow \infty$, then the test for $H_{0,\text{prop}}$*

$$\varphi_{n_1, n_2}^{\text{prop}} = \mathbf{1}\{T_{n_1, n_2} > \tilde{q}_{1-\alpha}\}$$

has asymptotic level α under $H_{0,\text{prop}}$ and asymptotic power 1 on the whole complement of $H_{0,\text{prop}}$. Here $\tilde{q}_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of

$$\begin{aligned} \mathcal{L} \left(\rho \left(\sqrt{\frac{n_1}{n}} \frac{\hat{W}_{n_2}^{(2)}}{\hat{A}_{n_1}^{(1)}} - \sqrt{\frac{n_2}{n}} \hat{W}_{n_1}^{(1)} \frac{\hat{A}_{n_2}^{(2)}}{[\hat{A}_{n_1}^{(1)}]^2}, \sqrt{\frac{n_1}{n}} \frac{\hat{W}_{n_2}^{(2)}(\tau)}{\hat{A}_{n_1}^{(1)}(\tau)} \right. \right. \\ \left. \left. - \sqrt{\frac{n_2}{n}} \hat{W}_{n_1}^{(1)}(\tau) \frac{\hat{A}_{n_2}^{(2)}(\tau)}{[\hat{A}_{n_1}^{(1)}(\tau)]^2} \right) \middle| \mathcal{C}_0 \right). \end{aligned}$$

Fig. 1 Illness-death model with recovery and transition hazards α_{01} , α_{10} , α_{02} , and α_{12} at time t

Table 1 Mean number of events per transition on [5, 30] provided by the simulation study of Sect. 5

Sample size	Transition			
	1 → 0	0 → 1	0 → 2	1 → 2
93	20.1	4.3	43.9	10.1
186	42.7	8.3	96.5	21.7
373	85.6	17.0	193.4	43.7
747	170.9	33.9	387.4	87.4
747 ^a	171	34	387	87

^aOriginal data

5 Simulation study

The motivating example behind the present simulation study is the SIR-3 data of Sect. 6. The setting is a specification of Example 1(a) called *illness-death model with recovery*. As illustrated in the multistate pattern of Fig. 1, the model has state space $\mathcal{S} = \{0, 1, 2\}$ and includes the transition hazards α_{01} , α_{10} , α_{02} , and α_{12} . The simulation of the underlying quantities is based on the methodology suggested by Allignol et al. (2011) generalized to the time-inhomogeneous Markovian multistate framework, which can be seen as a nested series of competing risks experiments. More precisely, the individual initial states are derived from the proportions of individuals at $t = 0$ and the censoring times are obtained from a multinomial experiment using probability masses equal to the increments of the censoring Kaplan–Meier estimate originated from the SIR-3 data. Similarly, event times are generated according to a multinomial distribution with probabilities given by the increments of the original Nelson–Aalen estimators. These times are subsequently included into the multistate simulation algorithm described in Beyersmann et al. (2012), Section 8.2. Since censoring times are sampled independently and each simulation step is only based on the current time and the current state, the resulting data follows a Markovian structure. A more formal justification of the multistate simulation algorithm can be found in Gill and Johansen (1990) and Theorem II.6.7 in Andersen et al. (1993).

We consider three different sample sizes: The original number of 747 patients is stepwisely reduced to 373, 186, and 93 patients. For each scenario we simulate 1000 studies. As an overview, the mean number of events for each possible transition and scenario is illustrated in Table 1.

The mean number of events regarding 747 patients reflects the original number of events. All numbers are restricted to the time interval [5,30], which is chosen due to

Fig. 2 95% confidence bands based on standard normal multipliers for the cumulative hazard of end-of-stay from the data example in Sect. 6. The solid black lines are the Nelson–Aalen estimators separately for ‘no ventilation’ (state 0, right plot) and ‘ventilation’ (state 1, left plot)

a small amount of events before $t = 5$ (left panel of Fig. 2). Further, less than 10% of all individuals are still under observation after day 30. In particular, asymptotic approximations tend to be poor at the left- and right-hand tails; cf. Remark 2(b) and Lin (1997).

Utilizing the R-package *sde* (Iacus 2014), the quantiles $c_{1-\alpha}^g$ in (4.3) of each single study are empirically estimated by simulating 1000 sample paths of a standard Brownian bridge. These quantiles are separately derived for both the Aalen- and Greenwood-type variance estimates (2.5) and (2.6). The bootstrap critical values are based on 1000 bootstrap realizations of $\hat{W}_n$ for each simulation step including both standard normal and centered Poisson variates with variance one. The latter is motivated by a slightly better performance compared to standard normal multipliers (Beyersmann et al. 2013; Dobler et al. 2017). Furthermore, Liu (1988) argued in a classical (linear regression) problem that wild bootstrap weights with skewness equal to one satisfy the second order correctness of the resampling approach. According to the cited simulation results, a similar result might hold true in our context, as the Poisson variates have skewness equal to one and standard normal variates are symmetric. A careful analysis of the convergence rates, however, is certainly beyond the scope of this article. In order to guarantee statistical reliability, we do not derive confidence bands for sample sizes and transitions with a mean number of observed transitions distinctly smaller than 20. The nominal level is set to $\alpha = 0.05$. All simulations are performed with the R-computing environment version 3.3.2 (R Core Team 2016).

Following Table 2, almost all bands constructed via Brownian bridges consistently tend to be rather conservative in our setting, i.e., result in too broad bands. Here, the usage of the Greenwood-type variance estimate yields more accurate coverage probabilities compared to the Aalen-type estimate. In contrast, the wild bootstrap approach mostly outperforms the Brownian bridge procedures: The log-transformed wild bootstrap bands approximately keep the nominal level even in the smaller sample sizes, except for the $0 \rightarrow 1$ transition with smallest sample size (corresponding to only 17 events in the mean; cf. Table 1). We also observe that the log-transformation

Table 2 Empirical coverage probabilities (%) from the simulation study of Sect. 5 separately for each transition and different simulated number of individuals

Transition	N	Type of confidence band					
		Brownian bridge			Wild bootstrap		
		95% log EP		95% log HW	95% log EP		95% log HW
		Aalen	Greenwood	Aalen	Greenwood	Poisson	Standard normal
0 → 1	373	96.4	95.9	95.5	95.3	92.5	92.5
	747	97.7	97.3	97.2	97.0	95.0	94.9
0 → 2	93	98.0	97.1	98.4	97.3	97.6	97.8
	186	98.3	95.5	98.9	97.4	97.2	98.2
1 → 0	373	98.1	95.0	98.2	96.9	97.2	97.3
	747	98.6	96.2	98.8	97.7	97.7	97.8
1 → 1	93	97.0	94.9	97.0	95.2	95.1	95.1
	186	97.3	95.8	97.7	96.1	95.6	95.7
1 → 2	373	97.2	96.3	97.9	97.0	95.2	95.3
	747	97.8	96.8	97.5	96.9	96.6	96.6
2 → 0	186	97.5	96.7	97.2	96.2	94.7	94.5
	373	98.2	97.7	98.2	97.8	95.8	95.8
2 → 1	747	97.2	96.6	96.6	96.0	94.4	95.5
2 → 2	93	97.2	96.6	96.6	96.0	94.4	95.5
	186	97.2	96.6	96.6	96.0	94.4	95.5
2 → 3	373	97.2	96.6	96.6	96.0	94.4	95.5
	747	97.2	96.6	96.6	96.0	94.4	95.5
3 → 0	93	97.2	96.6	96.6	96.0	94.4	95.5
	186	97.2	96.6	96.6	96.0	94.4	95.5
3 → 1	373	97.2	96.6	96.6	96.0	94.4	95.5
	747	97.2	96.6	96.6	96.0	94.4	95.5
3 → 2	93	97.2	96.6	96.6	96.0	94.4	95.5
	186	97.2	96.6	96.6	96.0	94.4	95.5
3 → 3	373	97.2	96.6	96.6	96.0	94.4	95.5
	747	97.2	96.6	96.6	96.0	94.4	95.5

EP: equal-precision band; HW: Hall–Wellner band

in general improves coverage for the wild bootstrap procedure. The current simulation study showed no clear preference for the choice of weight. Note that all wild bootstrap bands for transition $0 \rightarrow 2$ show a similar, but mostly reduced conservativeness compared to the bands provided by Brownian bridges. We have to emphasize that coverage probabilities for the cumulative hazard functions are drastically decreased to approximately 75% in all sample sizes if log-transformed pointwise confidence intervals would wrongly be interpreted time-simultaneously (results not shown).

The second set of simulations follows the test for proportional hazards derived in Theorem 4 with regard to keeping the preassigned error level under the null hypothesis. For that purpose, we assume a competing risks model with two competing events separately for two unpaired patient groups. For an illustration, see for instance, Figure 3.1 in Beyersmann et al. (2012).

We consider four different constant hazard scenarios: (I) the hazards for the type-1 event are set to $\alpha_{01}^{(1)}(t) = \alpha_{01}^{(2)}(t) = 2$ (no effect on the type-1 hazard, in particular, a hazard ratio of $c = 1$); (II) $\alpha_{01}^{(1)}(t) = 1$ and $\alpha_{01}^{(2)}(t) = 2$ (large effect); (III) $\alpha_{01}^{(1)}(t) = \alpha_{01}^{(2)}(t) = 1$; (IV) $\alpha_{01}^{(1)}(t) = 1$ and $\alpha_{01}^{(2)}(t) = 1.5$ (moderate effect). In each scenario, we set $\alpha_{02}^{(1)} = \alpha_{02}^{(2)}(t) = 2$, in particular, we consistently assume no group effect on the competing hazard. Further, scenario-specific administrative censoring times are chosen such that approximately 25% of the individuals are censored. The simulations designs are selected such that we include different effect sizes as well as different type-1 hazard ratio configurations with respect to the competing hazards. We consider a balanced design with $n_1 = n_2 = n \in \{125, 250, 500, 1000\}$. The right-hand tail of the domain of interest is set to $\tau = 0.3$. Simulation of the event times and types follows the procedure explained in Chapter 3.2 of Beyersmann et al. (2012). As before, we simulate 1000 studies for each scenario and sample size configuration, whereas the critical values of the Kolmogorov–Smirnov-type and Cramér-von-Mises-type statistics from Sect. 4.3 are derived from 1000 bootstrap samples utilizing both standard normal and centered Poisson variates with variance one.

The results for the type I error rates (for $\alpha = 0.05$) are displayed in Table 3. As expected from consistency, the type I error control becomes better with a larger number of patients for both test statistics in each scenario. Except for Scenario (II), all procedures keep the type I error rate quite accurately for $n \geq 500$. For smaller sample sizes, all tests tend to be conservative with a particular advantage for the Kolmogorov–Smirnov statistic.

6 Data example

The SIR-3 (Spread of Nosocomial Infections and Resistant Pathogens) cohort study at the Charité University Hospital in Berlin, Germany, prospectively collected data on the occurrence and consequences of hospital-acquired infections in intensive care (Beyersmann et al. 2006; Wolkewitz et al. 2008). A device of particular interest in critically ill patients is mechanical ventilation. The present data analysis investigates the impact of ventilation on the length of intensive care unit stay which is, e.g., of interest in cost-benefit analyses in hospital epidemiology (Beyersmann et al. 2011).

Table 3 Simulated size of $\phi_{n_1, n_2}^{\text{prop}}$ for nominal size $\alpha = 5\%$ under different sample sizes and constant hazard configurations

n_i	Scenario I			Scenario II			Scenario III			Scenario IV		
	KMS		CvM	KMS		CvM	KMS		CvM	KMS		CvM
	SN	Poi		SN	Poi		SN	Poi		SN	Poi	
125	0.029	0.024	0.033	0.029	0.026	0.027	0.023	0.045	0.041	0.028	0.030	0.042
250	0.035	0.039	0.039	0.040	0.038	0.037	0.034	0.039	0.040	0.037	0.034	0.034
500	0.057	0.054	0.059	0.034	0.038	0.040	0.041	0.056	0.053	0.047	0.045	0.044
1000	0.050	0.050	0.047	0.048	0.049	0.043	0.046	0.047	0.049	0.045	0.048	0.059

In each scenario $\tau = 0.3$ and 25% of individuals are censored

KMS Kolmogorov–Smirnov-type statistic, *CvM* Cramér–von-Mises-type statistic; *SN* standard normal multiplier; *Poi* centered poisson multiplier

Fig. 3 95% equal precision confidence bands based on standard normal multipliers and 95% log-transformed pointwise confidence intervals for the cumulative hazard of end-of-stay from the data example in Sect. 6. The solid black lines are the Nelson–Aalen estimators separately for ‘no ventilation’ (state 0, right plot) and ‘ventilation’ (state 1, left plot)

The analysis considers a random subset of 747 patients of the SIR-3 data which one of us has made publicly available (Beyersmann et al. 2012). Patients may either be ventilated (state 1 as in Fig. 1) or not ventilated (state 0) upon admission. Switches in device usage are modeled as transitions between the intermediate states 0 and 1. Patients move into state 2 upon discharge from the unit. The numbers of observed transitions are reported in the last row of Table 1. We start by separately considering the two cumulative end-of-stay hazards A_{12} and A_{02} , followed by a more formal group comparison as in Remark 2(c). Based on the approach suggested by Beyersmann et al. (2012), Section 11.3, we find it reasonable to assume the Markov property. Figure 2 displays the Nelson–Aalen estimates of A_{12} and A_{02} accompanied by simultaneous 95% confidence bands utilizing the 1000 wild bootstrap versions with standard normal variates and restricted to the time interval [5,30] of intensive care unit days. As before, the left-hand tail of the interval is chosen, because Nelson–Aalen estimation regarding A_{12} picks up at $t = 5$, cf. the left panel of Fig. 2. Graphical validation of empirical means and variances of $\hat{W}_n$ showed good compliance compared to the theoretical limit quantities stated in Remark 1. Bands using Poisson variates are similar (both results not shown). Figure 3 also displays the 95% pointwise confidence intervals based on a log-transformation. The performance of both equal precision and Hall–Wellner bands is comparable for transitions out of the ventilation state. However, the latter tend to be larger for the $0 \rightarrow 2$ transitions for later days due to more unstable weights at the right-hand tail. Equal precision bands are graphically competitive when compared to the pointwise confidence intervals. Ventilation significantly reduces the hazard of end-of-stay, since the upper half-space is not contained in the 95% confidence band of the cumulative hazard difference, see Fig. 4.

A second investigation exemplary applies the nonparametric proportionality test suggested in Sect. 4.3 to the present study example. For that purpose, we consider sex-specific subsamples (441 male vs. 306 female patients) and test $H_{0,\text{prop}}$ using both the Kolmogorov–Smirnov- and the Cramér-von-Mises-type test statistic. The

Fig. 4 95% confidence bands from relation (4.5) based on standard normal multipliers and 95% linear pointwise confidence intervals for difference of the two cumulative hazards of end-of-stay from the data example in Sect. 6. The solid black lines is the difference ‘ventilation vs. no ventilation’ of the Nelson–Aalen estimators within the two ventilation groups

Nelson–Aalen estimators separately displayed for males and females are given in Supplementary Figure S1. p values are computed from 1000 bootstrap samples utilizing standard normal variates and the right-hand limit of the interval of interest is set to 30 days. Tabulated results are in Supplementary Table S1 including the test statistics T_{n_1, n_2} and the bootstrap quantiles $\tilde{q}_{0.95}$ of Theorem 4 as well as the corresponding bootstrap p values $\tilde{p}$. Non-proportionality cannot be inferred for any of the transitions at the 5% level.

A concise R-script is available in the online supplementary material, which executes the bootstrap core operation given in relation (3.2) in a computationally efficient way.

7 Discussion and further research

We have given a rigorous presentation of a weak convergence result for the wild bootstrap methodology for the multivariate Nelson–Aalen estimator in a general setting only assuming Aalen’s multiplicative intensity structure of the underlying counting processes. This allowed the construction of time-simultaneous confidence bands and intervals as well as asymptotically valid equivalence and equality tests for cumulative hazard functions. In the context of time-to-event analysis, our general framework is not restricted to the standard survival or competing risks setting, but also covers arbitrary Markovian multistate models with finite state space, other classes of intensity models like relative survival or excess mortality models, and even specific semi-Markov situations. Additionally, independent left-truncation and right-censoring can be incor-

porated. The procedure has also been used to construct a test for proportional hazards. We want to emphasize that the framework induces a random number of multipliers in relation (3.2); thus, goes beyond existing approaches for competing risks as done in Beyersmann et al. (2013) or Lin (1997). Easy and computationally convenient implementation and within- or two-sample comparisons demonstrate its attractiveness in various practical applications.

Future work will be on the approximation of the asymptotic distribution corresponding to the matrix of transition probabilities (see Aalen and Johansen 1978) and functionals thereof in general Markovian multistate models. This is of great practical interest, because no similar Brownian Bridge procedure is available to perform time-simultaneous statistical inference. In particular, previous implications rely on pointwise considerations. Note that such an approach would significantly simplify the original justifications given by Lin (1997) and generalizes his idea mainly used in the context of competing risks (Scheike and Zhang 2003; Hyun et al. 2009; Beyersmann et al. 2013). In addition, we plan to extend the utilized wild bootstrap technique to general semiparametric regression models; see Lin et al. (2000) for an application in the survival context. Current work investigates to what degree the martingale properties presented in this article may be exploited to obtain wild bootstrap consistencies for such functionals of Nelson–Aalen estimates or for estimators in semiparametric regression models. We are confident that the present approach will lead to reliable inference procedures in these contexts for which there has been only little research on such general methodology.

In contrast to the procedure of Schoenfeld et al. (2002) or the framework in Cube et al. (2017), the more general illness-death model with recovery does not rely on a constant hazards assumption and captures both the time-dependent structure of mechanical ventilation and the competing event ‘death in ICU’. This significantly improves medical interpretations. The widths of the confidence bands were competitive compared to the pointwise confidence intervals, i.e., demonstrated usefulness in practical situations.

In the present data analysis, the multistate perspective is the natural way to assess the impact of time-dependent exposures on complex survival outcomes. This is a current topic in various fields of applied research; thus, the wild bootstrap procedure in its very general formulation is practicable to analyses regarding, for instance, different stages of illicit drug use (Mayet et al. 2012), the clinical course of liver diseases (Jepsen et al. 2015), antibiotics in hospital epidemiology (Munoz-Price et al. 2016), alternative outcomes in leukemia trials (Schmoor et al. 2013; Eefting et al. 2016), or joint replacements in orthopaedic patients (Gillam et al. 2012). It has even been recently applied in a study investigating femoral fracture risk, disability, and mortality in an elderly population (Bluhmki et al. 2017).

It has to be emphasized that our simulation study suggested that the wild bootstrap approach leads to more powerful procedures (i.e., to narrower confidence bands) compared to the approximation via Brownian bridges. As expected, the applied log-transformation results in improved small sample properties compared to the untransformed wild bootstrap bands. Based on the current simulation study, however, it was difficult to clearly recommend which type of band and which type of multiplier should be used.

Acknowledgements The authors would like to thank two referees and an associate editor for their helpful comments which increased the quality of this article.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

A Proofs

Proof (of Lemma 1) : Due to similarity, it is enough to concentrate at the first component only; thus, we subsequently suppress the subscript ‘1’. The independence of all white noise processes immediately imply the orthogonality of all component processes. At first, we verify the martingale property; the square-integrability is obviously fulfilled since $E(G_1^2(0)) < \infty$. To this end, let $0 \leq s \leq t$. By measurability of the counting and the predictable process with respect to $\mathcal{C}_0$, we have

$$\begin{aligned} E(\hat{W}_n(t) \mid \mathcal{C}_s) &= \sqrt{n} \int_{(0,t]} \frac{J(u)}{Y(u)} E(G(u) \mid \mathcal{C}_s) dN(u) \\ &= \sqrt{n} \int_{(0,s]} \frac{J(u)}{Y(u)} G(u) dN(u) \\ &\quad + \sqrt{n} \int_{(s,t]} \frac{J(u)}{Y(u)} E(G(u)) dN(u) = \hat{W}_n(s) \end{aligned}$$

by the independence of $\sigma(G(u))$ and $\mathcal{C}_s$ for all $u > s$. Hence, the martingale property is shown.

The predictable variation process $\langle \hat{W}_n \rangle$ is the compensator of $\hat{W}_n^2$, i.e., we calculate

$$\begin{aligned} E(\hat{W}_n^2(t) \mid \mathcal{C}_s) &= n \left(\int_{(0,s]} \int_{(0,s]} + \int_{(s,t]} \int_{(0,s]} + \int_{(0,s]} \int_{(s,t]} + \int_{(s,t]} \int_{(s,t]} \right) E(G(u)G(v) \mid \mathcal{C}_s) \\ &\quad \times \frac{J(u)J(v)}{Y(u)Y(v)} dN(u) dN(v) \\ &= n \left(\int_{(0,s]} \int_{(0,s]} G(u)G(v) + \int_{(s,t]} \int_{(0,s]} E(G(u))G(v) \right. \\ &\quad \left. + \int_{(0,s]} \int_{(s,t]} G(u)E(G(v)) \right. \\ &\quad \left. + \int_{(s,t]} \int_{(s,t]} E(G(u)G(v)) \right) \frac{J(u)J(v)}{Y(u)Y(v)} dN(u) dN(v) \\ &= n \left(\int_{(0,s]} \int_{(0,s]} G(u)G(v) \frac{J(u)J(v)}{Y(u)Y(v)} dN(u)dN(v) \right. \end{aligned}$$

$$\begin{aligned}
& + \int_{(s,t]} E(G^2(u)) \frac{J(u)}{Y^2(u)} dN(u) \Big) \\
& = \hat{W}_n^2(s) + n \int_{(0,t]} \frac{J(u)}{Y^2(u)} dN(u) - n \int_{(0,s]} \frac{J(u)}{Y^2(u)} dN(u),
\end{aligned}$$

again by the $\mathcal{C}_s$ -measurability of $G(u)$ for $u \leq s$ and their independence for $u > s$. The second to last equality is due to the independence of $G(u)$ and $G(v)$ for $u \neq v$. Hence, $(\hat{W}_n^2(t) - n \int_{(0,t]} \frac{J(u)}{Y^2(u)} dN(u))_{t \in [0,\tau]}$ is a martingale.

Letting Δf denote the jump-size process for a càdlàg function f , the definition of the optional variation process yields

$$\begin{aligned}
[\hat{W}_n](t) &= \sum_{0 < s \leq t} (\Delta \hat{W}_n(s))^2 = n \sum_{0 < s \leq t} G^2(u) \frac{J(u)}{Y^2(u)} \Delta N(u) \\
&= n \int_{(0,t]} G^2(u) \frac{J(u)}{Y^2(u)} dN(u),
\end{aligned}$$

where the sum is taken over all jump points of N . $\square$

Proof (of Theorem 2) : It is enough to verify the conditions of Rebolledo's martingale central limit theorem (in conditional probability); see e.g., Theorem II.5.1 in Andersen et al. (1993). Since the filtration $\mathcal{C}_0$ at time $s = 0$ is not trivial, the resulting weak convergence will hold given $\mathcal{C}_0$ as well, in probability. From the classical theory we know that the Aalen-type variance estimator, which is in fact the predictable variation process of $\hat{W}_n$, is uniformly consistent for the variance function.

It remains to prove the Lindeberg condition (2.5.3) on page 83 in Andersen et al. (1993). But, by the same arguments as in the proof of Lemma 1, this is exactly the same as the Lindeberg condition for the Nelson–Aalen estimator itself. And this holds due to the main assumption (2.2).

Hence, Rebolledo's martingale central limit theorem yields the desired weak convergence as well as the uniform consistency of the optional variation process. $\square$

Proof (of Theorem 3) : For convergence (4.1), see Section IV.1 in Andersen et al. (1993) in combination with Slutsky's theorem. Convergence (4.2) follows from the consistency of σ^{*2} , Slutsky's theorem and Theorem 2, since $\hat{W}_n$ asymptotically mimics the distribution of $\sqrt{n}(\hat{A}_n - A)$. The functional delta-method for $(x \mapsto \log x)$ completes the proof. $\square$

Proof (of Corollaries 1 and 3) : Due to the continuous limit distribution the conditional quantiles converge as well in probability; see e.g., Janssen and Pauls (2003), Lemma 1. The consistency of φ_n^{KS} under $K \neq$ follows from the convergence in probability of the conditional quantile towards a finite value and from the uniform consistency of the multivariate Nelson–Aalen estimator for the cumulative hazard functions. Since the factor $\sqrt{n}$ tends to infinity, the test statistic also goes to infinity in probability under $K \neq$. $\square$

Proof (of Corollary 2) : The proof extends the arguments of Wellek (2010), Section 3.1, from confidence intervals to confidence bands. Write $H = H_1 \cup H_2$ where

$$H_1 : \{A(s) \leq A_0(s) - \ell(s) \text{ for some } s \in [t_1, t_2]\}$$

and $H_2 : \{A(s) \geq A_0(s) + u(s) \text{ for some } s \in [t_1, t_2]\}.$

Suppose H is true and let without loss of generality be H_1 true due to analogy. Then the probability of a false rejection of H amounts to

$$\begin{aligned} &P(A_0(s) - \ell(s) < a_n(s) \text{ and } b_n(s) < A_0(s) + u(s) \text{ for all } s \in [t_1, t_2]) \\ &\leq P(A_0(s) - \ell(s) < a_n(s) \text{ for all } s \in [t_1, t_2]) \\ &\leq P(A(s) < a_n(s) \text{ for some } s \in [t_1, t_2]) \longrightarrow \alpha. \end{aligned}$$

Here the last inequality holds since H_1 is true and the convergence is due to the asymptotic coverage probability of the confidence band $(a_n(s), \infty)_{s \in [t_1, t_2]}$.

In order to prove consistency, suppose the alternative hypothesis K is true and choose any ε such that

$$0 < \varepsilon < \inf_{s \in [t_1, t_2]} -(A_0(s) - \ell(s) - A(s)) \wedge (A_0(s) + u(s) - A(s)).$$

Thus, by the (uniform) consistency of the Nelson–Aalen estimator and the wild bootstrap quantiles, the probability of a correct rejection of H equals

$$\begin{aligned} &P(A_0(s) - \ell(s) < a_n(s) \text{ and } b_n(s) < A_0(s) + u(s) \text{ for all } s \in [t_1, t_2]) \\ &\geq P(A(s) - \varepsilon < a_n(s) \text{ and } b_n(s) < A(s) + \varepsilon \text{ for all } s \in [t_1, t_2]) \longrightarrow 1 \end{aligned}$$

as $n \rightarrow \infty$. For the convergence in the previous display, also note that $a_n \xrightarrow{P} A$ as well as $b_n \xrightarrow{P} A$ uniformly in $[t_1, t_2]$. $\square$

Proof (of Theorem 4) :

Let $t_0 > 0$. Denote by $\mathfrak{D}_{>0}[t_0, \tau] \subset \mathfrak{D}[t_0, \tau]$ the cone of positive càdlàg functions that are bounded away from zero. It is easy to see that the functional $\phi : \mathfrak{D}_{>0}^2[t_0, \tau] \rightarrow \mathfrak{D}_{>0}[t_0, \tau]$, $(f, g) \mapsto \frac{f}{g}$ is Hadamard-differentiable tangentially to the set of pairs of continuous functions $\mathfrak{C}^2[t_0, \tau]$ with continuous and linear Hadamard-derivative

$$\phi'_{(f,g)} : \mathfrak{C}^2[t_0, \tau] \rightarrow \mathfrak{C}[t_0, \tau], \quad (h_1, h_2) \mapsto \frac{h_1}{g} - h_2 \frac{f}{g^2}.$$

A simpler Hadamard-differentiability result holds for ϕ 's restriction to τ , i.e., $\phi|_\tau : (0, \infty)^2 \ni (f(\tau), g(\tau)) \mapsto \frac{f(\tau)}{g(\tau)}$ with continuous, linear Hadamard-derivative

$$(\phi|_\tau)'_{(f,g)} : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad (h_1(\tau), h_2(\tau)) \mapsto \frac{h_1(\tau)}{g(\tau)} - h_2(\tau) \frac{f(\tau)}{g^2(\tau)}.$$

Hence, we apply the functional δ -method and the continuous mapping theorem to

$$\sqrt{\frac{n_1 n_2}{n}} \left(\phi(\hat{A}_{n_2}^{(2)}, \hat{A}_{n_1}^{(1)}) - \phi(A^{(2)}, A^{(1)}) \right) \quad \text{and} \quad \phi'_{(\hat{A}_{n_2}^{(2)}, \hat{A}_{n_1}^{(1)})} \left(\sqrt{\frac{n_1}{n}} \hat{W}_{n_2}^{(2)}, \sqrt{\frac{n_2}{n}} \hat{W}_{n_1}^{(1)} \right),$$

respectively, verifying their equality in distribution in the limit (conditionally in probability for the latter). Proceed similarly with the restricted functional $\phi|_{\tau}$. Furthermore, the difference functional of both above functionals ensures the Hadamard-differentiability tangentially to the set of pairs of continuous functions. Our specific choices of the distance ρ are continuous functionals. Hence we are able to apply the continuous mapping theorem again. To conclude the proof of the asymptotic behaviour of $\varphi_{n_1, n_2}^{\text{prop}}$ under $H_{0, \text{prop}}$, note that the particular weight function solves the problem of dividing by zero at $t_0 = 0$.

For the asymptotic power assertion, let $t_1 \in [0, \tau]$ at which $H_{0, \text{prop}}$ is violated. Then

$$\rho \left(\frac{\hat{A}_{n_2}^{(2)}}{\hat{A}_{n_1}^{(1)}}, \frac{\hat{A}_{n_2}^{(2)}(\tau)}{\hat{A}_{n_1}^{(1)}(\tau)} \right)$$

converges in probability to a positive value, whence $T_{n_1, n_2} \xrightarrow{P} \infty$ follows. The conditional quantiles, however, still converge to a finite constant in probability by the above arguments. $\square$

B Theoretical justification for the non-applicability of the ordinary multiplier resampling

In this part of the appendix, following the suggestion of a referee, we show in detail why the ordinary multiplier bootstrap fails to provide the correct limit process. For ease of presentation, we focus on the univariate Nelson–Aalen estimator in the special Markovian and uncensored multistate model: Let $(X(t))_t$ denote a Markov process with state space $\{1, 2, \dots, k\}$, where $k \geq 2$, i.e., given the present state of the process, the future development does not depend on previously occupied states. Without loss of generality, we focus on the transition from state 1 to state 2 in order to prove our claim. That is, we restrict attention to the cumulative hazard function given by $A(dt) = A_{12}(dt) = P(X(t+dt) = 2 \mid X(t) = 1)$; for ease of presentation, this subscript is omitted. Obviously, having shown that the competing resampling technique fails in this particular situation, the same is immediately implied in the general and multivariate case.

In this set-up, the alternative resampling approach is available because the counting process N for all $1 \rightarrow 2$ transitions is a sum of $n \in \mathbb{N}$ i.i.d. counting processes $N_i, i = 1, \dots, n$, i.e., $N = \sum_{i=1}^n N_i$. Hence, each counting process satisfies the

Doob-Meyer decomposition (2.1) such that we have the martingale representation

$$M(t) = N(t) - \int_{(0,t]} Y(s) dA(s).$$

The ordinary multiplier bootstrap works as follows: Consider the martingale representation of the Nelson–Aalen estimator,

$$W_n(t) = \sqrt{n} \int_{(0,t]} \frac{J(u)}{Y(u)} dM(u) + o_p(1) = \sqrt{n} \sum_{i=1}^n \int_{(0,t]} \frac{J(u)}{Y(u)} dM_i(u) + o_p(1),$$

where $o_p(1) \xrightarrow{P} 0$ as $n \rightarrow \infty$; cf. (2.3) and omit the index j therein. In the resampling step, the individual counting process martingales $M_i(u)$ are replaced with $G_i \cdot N_i(u)$. Here, $G_1, \dots, G_n$ are i.i.d. random variables with zero mean and unit variance, which are independent of the data. The resulting ordinary multiplier bootstrap process shall be denoted as

$$\tilde{W}_n(t) = \sqrt{n} \sum_{i=1}^n G_i \int_{(0,t]} \frac{J(u)}{Y(u)} dN_i(u).$$

We now show that this approach asymptotically does not yield the appropriate correlation structure. The limit processes of the Nelson–Aalen estimator and its wild bootstrap version as discussed in Section 3.2 are centered Gaussian martingales and, thus, have independent increments. On the other hand, this does not hold for the limit process of $\tilde{W}_n$: For any choice of $0 \leq s \leq t \leq \tau$ we have

$$\begin{aligned} & cov(\tilde{W}_n(t), \tilde{W}_n(s) \mid \mathcal{C}_0) \\ &= E(\tilde{W}_n(t) \tilde{W}_n(s) \mid \mathcal{C}_0) \\ &= n \sum_{i=1}^n \sum_{k=1}^n E(G_i G_k) \int_{(0,t]} \frac{J(u)}{Y(u)} dN_i(u) \int_{(0,s]} \frac{J(v)}{Y(v)} dN_k(v) \\ &= \frac{1}{n} \sum_{i=1}^n \int_{(0,t]} \frac{nJ(u)}{Y(u)} dN_i(u) \int_{(0,s]} \frac{nJ(v)}{Y(v)} dN_i(v) \\ &= \frac{1}{n} \sum_{i=1}^n \int_{(0,t]} \frac{1}{y(u)} dN_i(u) \int_{(0,s]} \frac{1}{y(v)} dN_i(v) + o_p(1) \\ &\xrightarrow{P} \tilde{\psi}(s, t) := E \left(\int_{(0,t]} \frac{1}{y(u)} dN_1(u) \int_{(0,s]} \frac{1}{y(v)} dN_1(v) \right), \end{aligned}$$

where the last convergence in probability holds due to the law of large numbers as $n \rightarrow \infty$. Let X denote the Markov process underlying the counting process N_1 . A twofold application of Fubini's theorem yields that the above expectation equals

$$\int_{(0,t]} \int_{(0,s]} \frac{1}{y(u)} \frac{1}{y(v)} E(dN_1(u)dN_1(v)).$$

As we assumed that there is no censoring, we have

$$E(dN_1(u)dN_1(v)) = P(X(u) = 1, X(u+du) = 2, X(v) = 1, X(v+dv) = 2).$$

Now, we distinguish between two different cases: First, we treat the diagonal which contributes the following integral to the asymptotic covariance:

$$\int_{(0,\min(s,t)]} \frac{E(dN_1(u))}{y^2(u)} = \int_{(0,\min(s,t)]} \frac{E(Y_1(u)dA(u))}{y^2(u)} = \int_{(0,\min(s,t)]} \frac{dA(u)}{y(u)},$$

where Y_1 is the at risk indicator of the process X for a transition out of state 1. This integral corresponds to the asymptotic covariance function of the Nelson–Aalen estimator, cf. Theorem 1. Unfortunately, the overall asymptotic covariance of the alternative multiplier approach is increased due to the off-diagonal integral parts:

$$\begin{aligned} & \int_{(0,t]} \int_{(0,s]} \frac{\mathbf{1}\{u \neq v\}}{y(u)y(v)} E(dN_1(u)dN_1(v)) \\ &= \int_{(0,s]} \int_{(0,s]} \frac{\mathbf{1}\{u \neq v\}}{y(u)y(v)} E(dN_1(u)dN_1(v)) \\ & \quad + \int_{(s,t]} \int_{(0,s]} \frac{\mathbf{1}\{u \neq v\}}{y(u)y(v)} E(dN_1(u)dN_1(v)) \\ &= 2 \int_{(0,s]} \int_{(u,s]} \frac{1}{y(u)y(v)} E(dN_1(u)dN_1(v)) \\ & \quad + \int_{(s,t]} \int_{(0,s]} \frac{1}{y(u)y(v)} E(dN_1(u)dN_1(v)) \end{aligned}$$

In the case of $u < v$ and due to the assumed Markov property, the above expected values reduce to

$$\begin{aligned} E(dN_1(u)dN_1(v)) &= P(X(u) = 1, X(u + du) = 2, X(v) = 1, X(v + dv) = 2) \\ &= P(X(v + dv) = 2 \mid X(v) = 1) P(X(v) = 1 \mid X(u + du) = 2) \\ &\quad \times P(X(u + du) = 2 \mid X(u) = 1) P(X(u) = 1) \\ &= P(X(u) = 1) P_{21}(u, v) dA(u) dA(v), \end{aligned}$$

where $P_{21}(u, v) = P(X(v) = 1 \mid X(u) = 2)$. All in all, using $y(u) = P(X(u) = 1)$, we obtain the following asymptotic covariance of the ordinary multiplier bootstrap version of the Nelson–Aalen estimator:

$$\tilde{\psi}(s, t) = \int_{(0, \min(s, t)]} \frac{dA(u)}{y(u)} + \int_{(0, s]} \int_{(0, t]} \frac{P_{21}(\min(u, v), \max(u, v))}{y(\max(u, v))} dA(u) dA(v).$$

We may conclude two things from this covariance representation: First, we see that $\tilde{\psi}(s, t) \neq \psi(s, t)$. That is, the asymptotic covariance function of the Nelson–Aalen estimator as given in Theorem 1 is altered by the ordinary multiplier bootstrap. As $\tilde{\psi}(s, t)$ is in general not a function of $\min(s, t)$, the underlying limit Gaussian process does not even have independent increments. Hence, the alternative multiplier approach fails in reproducing an adequately distributed stochastic process.

Second, we see that the correct limit covariance structure is ensured only if the second integral in the above representation of $\tilde{\psi}$ vanishes. This is the case if $P_{21} \equiv 0$ which holds, for example, if there is at most one $1 \rightarrow 2$ transition possible as in the competing risks special case. This is no surprise, though: In the competing risks set-up, both resampling options, our proposed wild bootstrap and the ordinary multiplier bootstrap, reduce to the same procedure; see Beyersmann et al. (2013) for the applicability of the wild bootstrap to Aalen–Johansen estimators in competing risks models.

References

- Aalen OO (1978) Nonparametric inference for a family of counting processes. *Ann Stat* 6(4):701–726
- Aalen OO, Johansen S (1978) An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat* 5(3):141–150
- Aalen OO, Borgan Ø, Gjessing HK (2008) *Survival and event history analysis: a process point of view*. Springer, New York
- Akritis MG (1986) Bootstrapping the Kaplan–Meier estimator. *J Am Stat Assoc* 81(396):1032–1038
- Allignol A, Schumacher M, Wanner C, Drechsler C, Beyersmann J (2011) Understanding competing risks: a simulation point of view. *BMC Med Res Methodol* 11:86
- Andersen PK, Væth M (1989) Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* 45(2):523–535
- Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) *Statistical models based on counting processes*. Springer, New York
- Bagdonavičius V, Levulienė R, Nikulin MS (2010) Goodness-of-fit criteria for the Cox model from left truncated and right censored data. *J Math Sci* 167(4):436–443
- Bajorunaite R, Klein JP (2007) Two-sample tests of the equality of two cumulative incidence functions. *Comput Stat Data Anal* 51(9):4269–4281

- Beyersmann J, Gastmeier P, Grundmann H, Bärwolff S, Geffers C, Behnke M, Rüden H, Schumacher M (2006) Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection. *Infect Control Hosp Epidemiol* 27(05):493–499
- Beyersmann J, Wolkewitz M, Allignol A, Grambauer N, Schumacher M (2011) Application of multistate models in hospital epidemiology: advances and challenges. *Biom J* 53(2):332–350
- Beyersmann J, Allignol A, Schumacher M (2012) *Competing risks and multistate models* with R. Springer, New York
- Beyersmann J, Di Termini S, Pauly M (2013) Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scand J Stat* 40(3):387–402
- Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. *Ann Stat* 9(6):1196–1217
- Bie O, Borgan Ø, Liestøl K (1987) confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scand J Stat* 14(3):221–233
- Bluhmki T, Peter RS, Rapp K, König H-H, Becker C, Lindlbauer I, Rothenbacher D, Beyersmann J, Büchele G (2017) Understanding mortality of femoral fractures following low-impact trauma in persons with and without care need. *J Am Med Dir Assoc* 18(3):221–226
- Chen Wei, Wang Dehui, Li Yanfeng (2015) A class of tests of proportional hazards assumption for left-truncated and right-censored data. *J Appl Stat* 42(11):2307–2320
- Cox D (1972) Regression models and life tables (with discussion). *J R Stat Soc* 34:187–220
- Curley MAQ, Wypij D, Watson RS, Grant MJC, Asaro LA, Cheifetz IM, Dodson BL, Franck LS, Gedeit RG, Angus DC, Matthay MA (2015) Protocolized sedation vs usual care in pediatric patients mechanically ventilated for acute respiratory failure: a randomized clinical trial. *J Am Med Assoc* 313(4):379–389 and for the RESTORE Study Investigators and the PALISI Network
- Dabrowska DM, Ho W-T (2000) Confidence bands for comparison of transition probabilities in a Markov chain model. *Lifetime Data Anal* 6(1):5–21
- de Wit M, Gennings C, Jenvey WI, Epstein SK (2008) Randomized trial comparing daily interruption of sedation and nursing-implemented sedation algorithm in medical intensive care unit patients. *Crit Care* 12(3):R70
- Dobler D, Pauly M (2014) Bootstrapping Aalen–Johansen processes for competing risks: handicaps, solutions, and limitations. *Electron J Stat* 8(2):2779–2803
- Dobler D, Beyersmann J, Pauly M (2017) Non-strange weird resampling for complex survival data. *Biometrika* 104(3):699–711
- Dudek A, Goćwin M, Leśkow J (2008) Simultaneous confidence bands for the integrated hazard function. *Comput Stat* 23(1):41–62
- Eefting M, de Wreede LC, Halkes CJM, Peter A, Kersting S, Marijt EWA, Veelken H, Putter H, Schetelig J, Falkenburg JHF (2016) Multi-state analysis illustrates treatment success after stem cell transplantation for acute myeloid leukemia followed by donor lymphocyte infusion. *Haematologica* 101(4):506–514
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* 7(1):1–26
- Efron B (1981) Censored data and the bootstrap. *J Am Stat Assoc* 76(374):312–319
- Gill R, Schumacher M (1987) A simple test of the proportional hazards assumption. *Biometrika* 74(2):289–300
- Gill RD, Johansen S (1990) A survey of product-integration with a view toward application in survival analysis. *Ann Stat* 18(4):1501–1555
- Gillam MH, Ryan P, Salter A, Graves SE (2012) Multi-state models and arthroplasty histories after unilateral total hip arthroplasties. *Acta Orthop* 83(3):220–226
- Glidden DV (2002) Robust inference for event probabilities with non-Markov event data. *Biometrics* 58(2):361–368
- Good PI (2005) *Permutation, parametric, and bootstrap tests of hypotheses*, 3rd edn. Springer, New York
- Grambsch PM, Therneau TM (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81(3):515–526
- Hall P, Wilson SR (1991) Two guidelines for bootstrap hypothesis testing. *Biometrics* 47(2):757–762
- Hall WJ, Wellner JA (1980) Confidence bands for a survival curve from censored data. *Biometrika* 67(1):133–143
- Hess KR (1995) Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat Med* 14(15):1707–1723

- Hieke S, Bertz H, Dettenkofer M, Schumacher M, Beyersmann J (2013) Initially fewer bloodstream infections for allogeneic vs. autologous stem-cell transplants in neutropenic patients. *Epidemiol Infect* 141(01):158–164
- Horvath L, Yandell BS (1987) Convergence rates for the bootstrapped product-limit process. *Ann Stat* 15(3):1155–1173
- Hyun S, Sun Y, Sundaram R (2009) Assessing cumulative incidence functions under the semiparametric additive risk model. *Stat Med* 28(22):2748–2768
- Iacus SM (2014) sde: simulation and inference for stochastic differential equations. <https://CRAN.R-project.org/package=sde>. R package version 2.0.13. Accessed 8 Nov 2017
- Janssen A, Pauls T (2003) How do bootstrap and permutation tests work? *Ann Stat* 31(3):768–806
- Jepsen P, Vilstrup H, Andersen PK (2015) The clinical course of cirrhosis: the importance of multistate models and competing risks analysis. *Hepatology* 62(1):292–302
- Kozioł JA, Byar DP (1975) Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data. *Technometrics* 17(4):507–510
- Kraus D (2007) Data-driven smooth tests of the proportional hazards assumption. *Lifetime Data Anal* 13(1):1–16
- Lin DY (1991) Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J Am Stat Assoc* 86(415):725–728
- Lin DY (1997) Non-parametric inference for cumulative incidence functions in competing risks studies. *Stat Med* 16(8):901–910
- Lin DY, Wei LJ, Ying Z (1993) Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 80(3):557–572
- Lin DY, Fleming TR, Wei LJ (1994) Confidence bands for survival curves under the proportional hazards model. *Biometrika* 81(1):73–81
- Lin DY, Wei LJ, Yang I, Ying Z (2000) Semiparametric regression for the mean and rate functions of recurrent events. *J R Stat Soc Ser B (Stat Methodol)* 62(4):711–730
- Liu L, Logan B, Klein JP (2008) Inference for current leukemia free survival. *Lifetime Data Anal* 14(4):432–446
- Liu RY (1988) Bootstrap procedures under some non-iid models. *Ann Stat* 16(4):1696–1708
- Lo SH, Singh K (1986) The product-limit estimator and the bootstrap: some asymptotic representations. *Probab Theory Relat Fields* 71(3):455–465
- Mayet A, Legleye S, Falissard B, Chau N (2012) Cannabis use stages as predictors of subsequent initiation with other illicit drugs among french adolescents: use of a multi-state model. *Addict Behav* 37(2):160–166
- Pauly M, Brunner E, Konietzschke F (2015) Asymptotic permutation tests in general factorial designs. *J R Stat Soc Ser B (Stat Methodol)* 77(2):461–473
- R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. Accessed 8 Nov 2017
- Sauaia A, Moore EE, Johnson JL, Ciesla DJ, Biffl WL (2009) Validation of postinjury multiple organ failure scores. *Shock* 31(5):438–447
- Scheike TH, Martinussen T (2004) On estimation and tests of time-varying effects in the proportional hazards model. *Scand J Stat* 31(1):51–62
- Scheike TH, Zhang M-J (2003) Extensions and applications of the Cox–Aalen survival model. *Biometrics* 59(4):1036–1045
- Schmoor C, Schumacher M, Finke J, Beyersmann J (2013) Competing risks and multistate models. *Clin Cancer Res* 19(1):12–21
- Schoenfeld DA, Bernard GR, Network ARDS (2002) Statistical evaluation of ventilator-free days as an efficacy measure in clinical trials of treatments for acute respiratory distress syndrome. *Crit Care Med* 30(8):1772–1777
- Schumacher M (1984) Two-sample tests of Cramér-von Mises- and Kolmogorov–Smirnov-type for randomly censored data. *Int Stat Rev* 52(3):263–281
- Silvia Munoz-Price L, Frencken Jos F, Sergey Tarima, Marc Bonten (2016) Handling time-dependent variables: antibiotics and antibiotic resistance. *Clin Infect Dis* 62(12):1558–1563. <https://doi.org/10.1093/cid/ciw191> ISSN 1537-6591
- Shu Y, Klein JP, Zhang M-J (2007) Asymptotic theory for the Cox semi-Markov illness-death model. *Lifetime Data Anal* 13(1):91–117

- Stewart RM, Park PK, Hunt JP, McIntyre RC Jr, McCarthy J, Zarzabal LA, Michalek JE; for the NIH/NHLBI ARDS Clinical Trials Network (2009) Less is more: improved outcomes in surgical patients with conservative fluid administration and central venous catheter monitoring. *J Am Coll Surg* 208(5):725–735
- Trof RJ, Beishuizen A, Cornet AD, de Wit RJ, Girbes ARJ, Groeneveld ABJ (2012) Volume-limited versus pressure-limited hemodynamic management in septic and nonseptic shock*. *Crit Care Med* 40(4):1177–1185
- von Cube M, Schumacher M, Wolkewitz M (2017) Basic parametric analysis for a multi-state model in hospital epidemiology. *BMC Med Res Methodol* 17(1):111
- Wellek S (2010) Testing statistical hypotheses of equivalence and noninferiority, 2nd edn. CRC Press, Boca Raton
- Whitt W (1980) Some useful functions for functional limit theorems. *Math Oper Res* 5(1):67–85
- Wolkewitz M, Vonberg RP, 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. *Crit Care* 12(2):R44
- Wu CFJ (1986) Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat* 14(4):1261–1295