


A Wild Bootstrap Approach for the Aalen–Johansen Estimator

Tobias Bluhmki ^{1,*} Claudia Schmoor,² Dennis Dobler,¹ Markus Pauly,¹ Juergen Finke,³
Martin Schumacher,⁴ and Jan Beyersmann¹

¹Institute of Statistics, Ulm University, Ulm, Germany

²Clinical Trials Unit, Medical Center Freiburg, University of Freiburg, Freiburg, Germany

³Department of Hematology, Oncology, and Stem-Cell Transplantation, Medical Center Freiburg,
University of Freiburg, Freiburg, Germany

⁴Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center,
University of Freiburg, Freiburg, Germany

**email*: tobias.bluhmki@uni-ulm.de

SUMMARY. We suggest a wild bootstrap resampling technique for nonparametric inference on transition probabilities in a general time-inhomogeneous Markov multistate model. We first approximate the limiting distribution of the Nelson–Aalen estimator by repeatedly generating standard normal wild bootstrap variates, while the data is kept fixed. Next, a transformation using a functional delta method argument is applied. The approach is conceptually easier than direct resampling for the transition probabilities. It is used to investigate a non-standard time-to-event outcome, currently being alive without immunosuppressive treatment, with data from a recent study of prophylactic treatment in allogeneic transplanted leukemia patients. Due to non-monotonic outcome probabilities in time, neither standard survival nor competing risks techniques apply, which highlights the need for the present methodology. Finite sample performance of time-simultaneous confidence bands for the outcome probabilities is assessed in an extensive simulation study motivated by the clinical trial data. Example code is provided in the web-based Supplementary Materials.

KEY WORDS: Blood cancer; Graft-versus-host-disease; Illness-death model; Resampling; Survival analysis; Time-dependent covariate.

1. Introduction

The present article is motivated by investigating the impact of graft-versus-host-disease (GvHD) prophylaxis on the probability to be alive without immunosuppressive therapy (IST) in leukemia patients who have undergone allogeneic haematopoietic cell transplantation. The challenge is that patients commonly require multiple IST episodes after transplantation, while at the same time always being at the risk of death. This type of data may occur in many clinical research questions whenever the interesting endpoint may appear in multiple episodes over time. Examples include the occurrence of adverse event or disease episodes after a specific therapy, or the time under a specific treatment, which can stop and start again.

Occurrence of GvHD is commonly analyzed via competing risks (e.g., Kalbfleisch and Prentice, 2002; Lazaryan et al., 2016); see also Pfirrmann et al. (2011) for general discussions of competing risks in leukemia trials. Here, the cumulative incidence function, that is, the cumulative probability of GvHD occurrence, is the probability parameter of interest. The aim of successful prophylaxis is, however, not solely the reduction of GvHD incidence, but also to concurrently increase the proportion of patients being alive and not immunosuppressively treated in the course of disease. Considering being alive without IST as a novel and non-standard time-to-event outcome has recently been suggested by, for

example, Schmoor et al. (2013) and Eefting et al. (2016). Here, a challenge is that the proportion of patients alive without IST is non-monotonic in time. Thus, competing risks techniques do not apply. Both Schmoor et al. (2013) and Eefting et al. (2016) suggested the use of multistate models (e.g., Andersen et al., 1993; Beyersmann et al., 2012) to jointly model changes of IST status while always being at the risk of death. However, a formal comparison of the outcome probabilities was not provided. The applied aim of the present article will be to investigate the impact of GvHD prophylaxis on the probability to be alive without IST using time-simultaneous confidence bands.

Such time-simultaneous inference is complicated by the fact that the asymptotic covariance function of the Aalen–Johansen estimator (Aalen and Johansen, 1978), the canonical nonparametric estimator of the matrix of transition probabilities, does not exhibit independent increments of the Gaussian limit process. To attack such situations, Lin and co-workers developed a simple and computationally convenient resampling procedure based on martingale representations (Lin et al., 1993, 1994). Here, one keeps the data fixed and replaces the unknown martingale increments using standard normal variates. Next, the distribution of the estimator is approximated by repeatedly generating such variates, see Martinussen and Scheike (2006) for a textbook treatment. In a nutshell, the idea is to replace asymptotic normality by finite

sample normality with approximately the right covariance, while basing the resampling on martingale representations does not require the strict independent and identically distributed (iid) setup of the standard bootstrap. The approach is an example of the more general “wild bootstrap,” which was originally developed for heteroscedastic regression analysis (e.g., Liu, 1988; Mammen, 1992; Davidson and Flachaire, 2008). Here, independent random variates with expectation zero and variance one—called multipliers—are used.

Lin (1997) applied the wild bootstrap to construct time-simultaneous confidence bands for the cumulative incidence function from right-censored competing risks data. Beyersmann et al. (2013) as well as Dobler et al. (2017) recently gave a thorough mathematical treatment of this approach also allowing for independent left-truncation and general multipliers. Extensions to more complex multistate models are, however, rare in the current literature. The major challenges are involved martingale representations of the Aalen–Johansen estimator (Andersen et al., 1993, p. 320). Thus, we first apply wild bootstrap resampling to approximate the distribution of the multivariate standardized Nelson–Aalen estimator of the cumulative transition hazards. Using a functional delta method-type argument, we conclude that the wild bootstrap approximation for the Nelson–Aalen estimator, transformed according to the Hadamard derivative of the product integral (Gill and Johansen, 1990), mimics the limit distribution of the Aalen–Johansen estimator. Our approach considerably simplifies the original arguments given by Lin (1997) and Beyersmann et al. (2013) for competing risks and generalizes to arbitrary time-inhomogeneous Markov processes with finite state space subject to independent left-truncation and right-censoring. This general formulation allows for reversible multistate models, such as the so-called illness-death model with (IST-) recovery applied in Schmoor et al. (2013), as well as irreversible models; see Liu et al. (2008) for an example in the context of current leukemia-free survival. Allowing for left-truncation is motivated by observational studies on, for instance, pregnancy outcomes (Beyersmann et al., 2012).

The present article is organized as follows: multistate models and the connection between cumulative transition hazards and transition probabilities are briefly presented in Section 2. Wild bootstrap resampling for the multivariate Nelson–Aalen estimator and the main result regarding the transformation onto the Aalen–Johansen estimator is introduced in Section 3. Section 4 considers constructing simultaneous confidence bands using the present resampling approach. A simulation study investigating coverage probabilities of such bands is reported in Section 5, where the simulation setup is closely motivated by the leukemia data of Schmoor et al. (2013), which is re-analyzed in Section 6. A discussion is offered in Section 7. Technical proofs as well as example R code applied to freely accessible data are deferred to the web-based Supplementary Material.

2. Multistate Models

Let $(X_t)_{t \geq 0}$ be a time-inhomogeneous Markov process with state space $\mathcal{S} = \{0, 1, 2, \dots, \mathcal{J}\}$, $\mathcal{J} \in \mathbb{N}$, and càdlàg sample paths, that is, right-continuous with left-hand limits. Consider

the transition probability matrix $\mathbf{P}(s, t) = \{P_{lj}(s, t)\}_{l, j \in \mathcal{S}}$ with transition probabilities

$$P_{lj}(s, t) = P(X_t = j | X_s = l) = P(X_t = j | X_s = l, \text{Past}), \quad s \leq t, \quad l, j \in \mathcal{S}.$$

We assume the existence of transition hazards $\alpha_{lj}(t)$ fulfilling

$$\alpha_{lj}(t) \cdot dt = P(X_{t+dt} = j | X_{t-} = l), \quad l, j \in \mathcal{S}, \quad l \neq j$$

and define $\alpha_{ll}(t) = -\sum_{j=0, j \neq l}^{\mathcal{J}} \alpha_{lj}(t)$ for all $l \in \mathcal{S}$. The connection between transition hazards and transition probabilities is established through product integration (Aalen and Johansen,

1978): Writing $\mathbf{A}(t) = \{A_{lj}(t)\}_{l, j \in \mathcal{S}} = \left\{ \int_0^t \alpha_{lj}(u) du \right\}_{l, j \in \mathcal{S}}$,

transition probability matrix can be expressed as a product integral

$$\mathbf{P}(s, t) = \prod_{u \in (s, t]} \{ \mathbf{I} + d\mathbf{A}(u) \}, \quad (1)$$

where \mathbf{I} is the $(\mathcal{J} + 1) \times (\mathcal{J} + 1)$ identity matrix. The right-hand side of (1) is the limit of $\prod_{k=1}^K \{ \mathbf{I} + \Delta \mathbf{A}(t_k) \}$ for increasingly finer partitions $s = t_0 < t_1 < \dots < t_{K-1} < t_K = t$, $K \in \mathbb{N}$, where the (l, j) -th entry of $\Delta \mathbf{A}(t_k)$ is defined by $A_{lj}(t_k) - A_{lj}(t_{k-1})$.

3. Wild Bootstrapping the Aalen–Johansen Estimator

Let n be the number of individuals under study, where the individual trajectories are conditionally independent replicates of the process $(X_t)_{t \geq 0}$ given the states occupied at time origin. Assume that observations are subject to independent left-truncation and right-censoring (see Andersen et al., 1993, Chapter III for details), which may be state-dependent. Applying the usual counting process notation, the process $N_{i;l;j}(t)$ counts the number of observed direct transitions of individual i , $i = 1, \dots, n$, from state l into state j in the time interval $[0, t]$. In contrast to the simpler standard survival and competing risks settings, $N_{i;l;j}(t)$ may have values greater than one in general multistate problems. Let

$$Y_{i;l}(t) = \mathbf{1}(\text{individual } i \text{ is observed to be in state } l \text{ just prior to } t)$$

denote the at risk indicator for an observed (l, j) -transition of individual i just prior to t . Aggregation over all individuals yields $N_{lj}(t) = \sum_{i=1}^n N_{i;l;j}(t)$ and $Y_l(t) = \sum_{i=1}^n Y_{i;l}(t)$. Then, the Nelson–Aalen estimator $\hat{\mathbf{A}}(t)$ of $\mathbf{A}(t)$ has (l, j) -th entry, $l \neq j$,

$$\hat{A}_{lj}(t) = \int_0^t \mathbf{1}\{Y_l(u) > 0\} \frac{dN_{lj}(u)}{Y_l(u)} = \sum_{i=1}^n \int_0^t \mathbf{1}\{Y_{i;l}(u) > 0\} \frac{dN_{i;l;j}(u)}{Y_{i;l}(u)}.$$

The diagonal entries of $\hat{\mathbf{A}}(t)$ are such that the sum of each row equals 0. We note that $\hat{A}_{lj}(t)$ is a finite sum with increments at the observed transition times from state l to j .

Under standard regularity assumptions, Theorem IV.1.2 in Andersen et al. (1993) provides convergence in distribution

on compact time intervals

$$\mathbf{W} = \sqrt{n} \cdot (\widehat{\mathbf{A}} - \mathbf{A}) \xrightarrow{\mathcal{D}} \mathbf{U} = (U_{ij})_{i,j \in \mathcal{S}}, \quad (2)$$

where the non-diagonal entries of \mathbf{U} are independent Gaussian martingales with $U_{ij}(0) = 0$ and almost surely continuous sample paths. The diagonal entries of \mathbf{U} are also such that the sum of each row equals 0. Convergence in distribution “ $\xrightarrow{\mathcal{D}}$ ” is considered for $n \rightarrow \infty$ on the matrix-valued càdlàg function space endowed with the product Skorohod topology. Weak convergence in (2) is proven using martingale arguments, because it holds that

$$\sqrt{n} \left\{ \widehat{A}_{lj}(t) - A_{lj}(t) \right\} - \sqrt{n} \left[\int_0^t \frac{\mathbf{1}\{Y_l(u) > 0\}}{Y_l(u)} d \left\{ \sum_{i=1}^n M_{i;l j}(u) \right\} \right] \xrightarrow{\mathcal{L}} 0, \quad (3)$$

with transition- and individual-specific martingales $M_{i;l j}(t) = N_{i;l j}(t) - \int_0^t \alpha_{lj}(u) Y_{i;l}(u) du$. We now use the martingale representation (3) to introduce the resampling procedure.

In order to approximate the Gaussian limit process \mathbf{U} in (2), we extend the resampling scheme proposed by Lin (1997) to the multivariate Nelson–Aalen estimator. The idea is to keep the data fixed and substitute the unknown martingale quantities $dM_{i;l j}(t)$ with $dN_{i;l j}(t)$ times a standard normal random variable. “Keeping the data fixed” means that all derivations are conditioned on the available data. To be concrete, consider a $(\mathcal{J} + 1) \times (\mathcal{J} + 1)$ matrix-valued process $\boldsymbol{\xi}$ with non-diagonal entries,

$$\xi_{lj}(t) = \sqrt{n} \cdot \sum_{i=1}^n \int_0^t G_{i;l j}(u) \cdot \mathbf{1}\{Y_l(u) > 0\} \frac{dN_{i;l j}(u)}{Y_l(u)}, \quad l \neq j, \quad (4)$$

which can be derived from the martingale representation of the Nelson–Aalen estimator (3) by introducing iid standard normal variables $G_{i;l j}(u)$. This is the typical choice in biometrical applications; however, the theory even allows for more general multipliers with expectation 0 and variance 1. The diagonal entries of $\boldsymbol{\xi}$ are again such that the sum of each row equals 0. Note that (4) requires a random multiplier for each observed transition time of each individual. The resampling in (4) extends the wild bootstrap of Lin (1997) and Beyersmann et al. (2013) to all single entries of the multivariate Nelson–Aalen estimator $\widehat{\mathbf{A}}$, allowing for repeated $l \rightarrow m$ transitions of an individual. As a result, $\boldsymbol{\xi}$ asymptotically mimics the distribution \mathbf{U} in (2) given the data, see Web Appendix S1 for details. Mathematically, the wild bootstrap is not required at this stage, because the process \mathbf{U} in (2) has independent increments. However, this property gets lost when transforming $\boldsymbol{\xi}$ for making inference on the transition probabilities. Plugging the Nelson–Aalen estimator $\widehat{\mathbf{A}}$ into (1), the transition matrix \mathbf{P} can be estimated by means of the Aalen–Johansen estimator

(Aalen and Johansen, 1978)

$$\widehat{\mathbf{P}}(s, t) = \prod_{u \in (s, t]} \left\{ \mathbf{I} + d\widehat{\mathbf{A}}(u) \right\},$$

which is a finite matrix product over all event times in $(s, t]$, $s < t$. Applying a functional delta method-type argument, Theorem IV.4.2 in Andersen et al. (1993) states convergence in distribution on compact time intervals,

$$\mathbf{B}(s, \cdot) = \sqrt{n} \cdot \left\{ \widehat{\mathbf{P}}(s, \cdot) - \mathbf{P}(s, \cdot) \right\} \xrightarrow{\mathcal{D}} \int_s^\cdot \mathbf{P}(s, u) d\mathbf{U}(u) \mathbf{P}(u, \cdot), \quad (5)$$

where \mathbf{U} is as in (2). At this stage, resampling is the method of choice, because the asymptotic Gaussian process lacks independent increments. However, martingale representations for (5) are much more involved compared to (3), see Andersen et al. (1993), equation (4.4.7) for general multistate models, and this is even true for the special competing risks model (Lin, 1997). To overcome this issue, Web Appendix S1 proves that, given the data,

$$\boldsymbol{\zeta}(s, \cdot) = \int_s^\cdot \widehat{\mathbf{P}}(s, u) d\boldsymbol{\xi}(u) \widehat{\mathbf{P}}(u, \cdot) \quad (6)$$

possesses the same limit behavior as $\mathbf{B}(s, \cdot)$. Consequently, (5) can be approximated in two subsequent steps: First, we resample on the hazard scale by means of generating a large number of replicates $\boldsymbol{\xi}$, and second, we transform them via (6) according to the Hadamard derivative of the original hazards-to-probabilities functional in order to obtain probability statements. This is not unlike a Bayesian approach, but explicitly builds on functional delta method asymptotics. Note that quantity (6) includes the resampling of Lin (1997) as a special case. Unlike resampling with replacement from the individual trajectories under an iid setup (Efron, 1981), resampling based on (6) works in the more general martingale setup outlined above, see Andersen et al. (1993), Section IV.1.4. for an in-depth discussion. Further, the present resampling scheme estimates n different distributions by means of only n individuals, which is why Mammen (1992) called such a procedure “wild.” We also emphasize the close link between relation (4) and general wild bootstrap representations considered in, for instance, Pauly (2011). Note that the procedure works due to binary increments $dN_{i;l j}(u)$, which is a natural assumption, because the entire theory is developed under a time-continuous framework (Andersen et al., 1993) implying that at most one individual $l \rightarrow j$ transition is observed at time t .

The two key-steps of the proof are as follows: in the spirit of the functional delta transformation in (6), the Continuous Mapping Theorem first verifies that

$$\int_s^\cdot \mathbf{P}(s, u) d\boldsymbol{\xi}(u) \mathbf{P}(u, \cdot) \xrightarrow{\mathcal{D}} \int_s^\cdot \mathbf{P}(s, u) d\mathbf{U}(u) \mathbf{P}(u, \cdot),$$

in probability, that is, given the data. Then, uniform consistency of the Aalen–Johansen estimator in combination with Lenglar’s inequality proves

$$\sup_{t \in [s, \tau]} \left| \int_s^t \widehat{\mathbf{P}}(s, u) d\widehat{\boldsymbol{\xi}}(u) \widehat{\mathbf{P}}(u, t) - \int_s^t \mathbf{P}(s, u) d\boldsymbol{\xi}(u) \mathbf{P}(u, t) \right| \xrightarrow{P} 0 \text{ in probability.}$$

Here, $|\cdot|$ denotes the Euclidean norm for matrices. The theoretical requirement for the correct wild bootstrap approximation on $[s, \tau]$ is that the asymptotic probabilities of non-empty risk sets are bounded away from zero on that interval of interest. This is similar to standard survival studies, where τ is commonly chosen as the largest observed event time.

4. Simultaneous Confidence Bands

The aim of the present section is to utilize the wild bootstrap derived in Section 3 for the convenient construction of asymptotic confidence bands for the outcome probabilities of interest and statistical two-sample tests for independent samples. For that purpose, fix time s , and define $\widehat{P}_{lj}(s, t)$ and $\zeta_{lj}(s, t)$ as the (l, j) -th entries of the corresponding matrix-valued processes $\widehat{\mathbf{P}}(s, t)$ and $\boldsymbol{\zeta}(s, t)$, $l, j \in \mathcal{S}$, $s < t$. Introduce

$$C_{lj}(s, t) = \sqrt{n} \cdot g(t) \left[\phi \left\{ \widehat{P}_{lj}(s, t) \right\} - \phi \left\{ P_{lj}(s, t) \right\} \right]$$

as the weighted and transformed variant of the (l, j) -th entry of $\mathbf{B}(s, t)$, where $g(\cdot)$ is a weight function and $\phi(\cdot)$ a transformation with non-zero continuous derivative $d\phi(\cdot)$. Different weightings lead to different types of confidence bands, whereas the rationale of transformations is to improve small sample performance. As suggested in Lin (1997), we focus on the log-log transformation $\phi(x) = \log\{-\log(1-x)\}$, $x \in (0, 1)$ to ensure that the confidence bands are contained in $[0, 1]$. Consider the weight function

$$g(t) = \frac{\{\widehat{P}_{lj}(s, t) - 1\} \cdot \log\{1 - \widehat{P}_{lj}(s, t)\}}{\sqrt{n \cdot \widehat{\text{var}}\{\widehat{P}_{lj}(s, t)\}}},$$

with $t \geq s$ and $\widehat{\text{var}}\{\widehat{P}_{lj}(s, t)\}$ denoting the empirical variance of the wild bootstrap realizations of $\zeta_{lj}(s, t)$ divided by n . Following Chapter IV.3.3 in Andersen et al. (1993), the resulting bands for the transition probability are called equal-precision (EP) bands (cf. also the choice in Lin, 1997).

Applying a functional delta method-type argument, the conditional distribution of

$$\widehat{C}_{lj}(s, t) = g(t) \cdot d\phi\{\widehat{P}_{lj}(s, t)\} \cdot \zeta_{lj}(s, t) \quad (7)$$

can be utilized to approximate the distribution of $C_{lj}(t)$. Let q_α be the conditional $(1 - \alpha)$ quantile such that, given the data,

$$P \left\{ \sup_{t \in [t_1, t_2]} |\widehat{C}_{lj}(s, t)| > q_\alpha \right\} = \alpha, \quad (8)$$

where $\alpha \in (0, 1)$ and $s \leq t_1 < t_2 \leq \tau$; see Section 6 for an example of choosing t_1 and t_2 in practice. Then, an asymptotic $(1 - \alpha)$ confidence band for $\phi \left\{ P_{lj}(s, t) \right\}$ is given by

$$\phi \left\{ \widehat{P}_{lj}(s, t) \right\} \pm \frac{q_\alpha}{\sqrt{n} \cdot g(t)}, \quad t \in [t_1, t_2]. \quad (9)$$

Applying the inverse function $\phi^{-1}(\cdot)$, a confidence band for $P_{lj}(s, t)$ can be derived.

The present approach can easily be extended to construct confidence bands for the difference $P_{lj}^{(1)} - P_{lj}^{(2)}$ of two transition probabilities from two independent samples. In the analysis of leukemia patients below, we will be interested in the difference of the probabilities to be alive without IST between prophylaxis groups. Let

$$D(s, t) = k(t) \left[\widehat{P}_{lj}^{(1)}(s, t) - \widehat{P}_{lj}^{(2)}(s, t) - \left\{ P_{lj}^{(1)}(s, t) - P_{lj}^{(2)}(s, t) \right\} \right], \quad (10)$$

where $k(t)$ is another positive weight function. The distribution of quantity (10) can be approximated by

$$\widehat{D}(s, t) = k(t) \left\{ \frac{\zeta_{n_1;lj}(s, t)}{\sqrt{n_1}} - \frac{\zeta_{n_2;lj}(s, t)}{\sqrt{n_2}} \right\}, \quad (11)$$

where $\zeta_{n_r;lj}(s, t)$ are the wild bootstrap versions of $\sqrt{n_r} \cdot \left\{ \widehat{P}_{lj}^{(r)}(s, t) - P_{lj}^{(r)}(s, t) \right\}$, $r = 1, 2$, and sample sizes n_r in group r . For our purposes, we choose $k \equiv 1$. This choice yields the approximate $(1 - \alpha)$ confidence band for the difference of the two transition probabilities given by

$$\left\{ \widehat{P}_{lj}^{(1)}(s, t) - \widehat{P}_{lj}^{(2)}(s, t) \right\} \pm \tilde{q}_\alpha, \quad (12)$$

where the quantile \tilde{q}_α is approximated such that $P \left\{ \sup_{t_1 \leq t \leq t_2} |\widehat{D}(s, t)| > \tilde{q}_\alpha \right\} = \alpha$. The confidence band for the difference can also be viewed as a Kolmogorov–Smirnov-type asymptotic level α test.

In practice, the wild bootstrap for the Aalen–Johansen estimator can be realized in the statistical software **R** as follows: the data file is assumed to be arranged in “long” format (Beyersmann et al., 2012), that is, each row represents one individual “at risk” for a certain transition. The columns indicate transition-type, censoring status, and when the patient started and stopped being at risk for a certain transition. We then implement the following algorithm:

1. For each observed event time u and each $l \rightarrow j$ transition
 - compute the number of observed $l \rightarrow j$ transitions at u and the number of individuals in state l just prior to u .
 - generate as many $G_{i;lj}(u)$ as $l \rightarrow j$ transitions are observed at u and compute $\widehat{P}_{lj}(s, u)$ and $\widehat{P}_{lj}(u, \cdot)$.

2. Compute ξ using (4) and arrange the increments $d\xi$ in a $(\mathcal{J} + 1) \times (\mathcal{J} + 1)$ matrix for each u .
3. Compute ζ via (6), which is used to derive one replicate of $\widehat{C}_{lj}(s, t)$ in (7).
4. Repeat steps (1)–(3) m times, where m is the number of bootstrap iterations (say 1000). The empirical $(1 - \alpha)$ quantile of the m replicates $\sup_{t \in [t_1, t_2]} |\widehat{C}_{lj}(s, t)|$ yields an approximation of the desired quantile q_α in (8).

The R-package `etm` (Allignol et al., 2011) provides fast computation of the Aalen–Johansen estimator \widehat{P} . Example R-code applying the algorithm to a freely accessible dataset is available as Supplementary Material, cf. Web Appendix S2 for explanations.

5. Simulation Study

The present simulation study uses the published data of Schmoor et al. (2013), which motivated the methodological developments of this article, as a template. The setting is an illness-death model with recovery as illustrated in Figure 1. The data is re-analyzed in Section 6. Data generation follows the simulation technique suggested by Fiocco et al. (2008) and Allignol et al. (2011). More precisely, the increments of the Nelson–Aalen estimators computed from the published data are utilized to determine the transition time and type by means of the multistate simulation algorithm described in detail in Beyersmann et al. (2012), Section 8.2. The idea is to generate multistate trajectories as a sequence of competing risks experiments. Random right-censoring follows a multinomial experiment with probabilities equal to the increments of the observed censoring Kaplan–Meier estimator (Web Figure S3). In analogy to the original data, all individuals are assumed to start in state 1 of being alive and under IST. More details regarding data generation can be found in Web Appendix S3.

We consider five different scenarios: the original sample size of 103 patients in the treatment group is increased to 200, 300, 400, and 500 patients. The focus is on $P_{10}(0, t)$, which is one of the relevant quantities of interest within the real data analysis of Section 6. For each scenario, we consider 1000 simulated studies, and the empirical approximation of the quantile q_α within each study is based on 1000 realizations of ζ using standard normal multipliers. We focus on log-log transformed EP bands restricted to the intervals $[t_1, t_2] \in \{[4, 12], [4, 24], [4, 48]\}$ interpreted as time in months in order to investigate different interval widths. The left limit

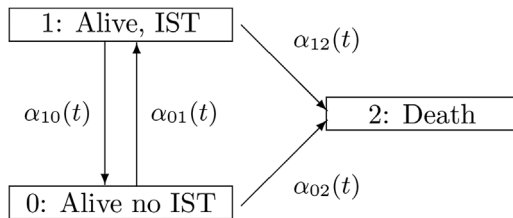


Figure 1. Illness-death model with recovery based on the randomized clinical trial data example with transition-specific hazards $\alpha_{01}(t)$, $\alpha_{02}(t)$, $\alpha_{10}(t)$, and $\alpha_{12}(t)$. Patients start in state 1 of being under IST.

Table 1

Coverage probabilities of log-log EP confidence bands (CBs) regarding $P_{10}(0, t)$ (1000 bootstrap iterations with standard normally distributed multipliers) separately computed for each sample size and $[t_1, t_2]$ considered in the simulation study of Section 5. The confidence level is set to 95%.

n	Coverage in % on $[t_1, t_2]$		
	$[4, 12]$	$[4, 24]$	$[4, 48]$
103	93.4	92.7	92.9
200	94.4	94.3	93.8
300	93.6	93.5	94.0
400	95.4	94.7	93.7
500	94.9	94.3	94.2

of the intervals equals 4, because IST is mandatory for around the first three months after transplantation and, consequently, no $1 \rightarrow 0$ transitions are available beforehand. The confidence level is set to 0.95.

According to Table 1, coverages are slightly too low for the smallest sample size of 103 patients ($\approx 93\%$). Concerning moderate and large sample sizes, the empirical coverage probabilities approach the confidence level on all time intervals. We also observe that a shorter time interval of interest leads to an increased performance of the wild bootstrap confidence bands. The intuition is that a smaller time period of interest leads to a better coverage, since less statistical uncertainty has to be captured.

We also perform three additional simulation studies (Web Appendix S4): First, the construction of confidence bands is based on centered Poisson multipliers with variance one. They are motivated by the fact that they possibly mimic the counting process data structure more closely, which has led to a slightly improved performance in the more simple competing risks setting (Beyersmann et al., 2013; Dobler and Pauly, 2014). Following recent findings in Dobler et al. (2017), heuristic arguments suggest that second order correctness may be achieved by using multipliers with unit skewness, which does not hold for the standard normal choice. However, the present results showed no clear preference for using Poisson multipliers throughout all scenarios. Second, we investigate the robustness of the procedure regarding the number of bootstrap iterations. For that purpose, we change the original number of iterations in both directions. It is shown that deviations are negligible compared to the current results when 2000 replicates are used, whereas a reduction to only 500 replicates results in generally lower coverages particularly for the broadest (and most challenging) time interval $[4, 48]$. A third study demonstrates the validity of the wild bootstrap in the presence of two different external random left-truncation mechanisms. However, a larger sample size is needed in order to obtain comparable results as without left-truncation. The intuitive reason is that less information is available due to delayed study entry, where some individuals never enter the study because of a terminal event before potential study entry. Those individuals under study, however, often do not contribute to early risk sets.

6. Being Alive and Without IST as an Outcome in Leukemia Trials

We now re-analyse the study data on allogeneic haematopoietic cell transplantation in leukemia patients (e.g., Socié et al., 2011; Schmoor et al., 2013) that motivated our methodological developments. In this context, GvHD after transplantation is a severe side effect inducing increased morbidity and mortality. The original study hypothesized that Anti-T-cell globulins decrease the incidence of GvHD. Studies on GvHD prophylaxis are often based on a competing risks framework (see e.g., Kalbfleisch and Prentice, 2002; Lazaryan et al., 2016); however, an additional aim of successful GvHD prophylaxis is to increase the proportion of patients that are both alive and do not require IST. Therefore, Socié et al. (2011) compared the impact of standard GvHD prophylaxis with and without pretransplantation Grafalon (formerly ATG-Fresenius S = ATG-F) medication on the time under IST in 201 randomly assigned patients (Grafalon $n = 103$, control $n = 98$). Since multiple episodes of IST are commonly observed during follow-up, an illness-death model with recovery was applied accounting for the time-dependent nature of IST. Besides death and censoring, this setting allows patients to switch back and forth between “IST” and “no IST.” A

multistate pattern is given in Figure 1. All transplants were allogeneic from unrelated donors; consequently, all patients required IST for around 3 months after transplantation. Cox proportional hazards models showed a significant decrease in the hazard of being alive and under IST in Grafalon treated patients. The result was graphically supported by the comparison of the corresponding Aalen–Johansen estimators (Schmoor et al., 2013). The latter estimates the non-monotonic probabilities in time of being alive and either under or free of IST; however, a formal comparison of the probabilities of interest was not provided. The present article completes this approach using confidence bands for a two-group comparison of the outcome probabilities of interest.

The construction of confidence bands is based on 1000 realizations of ζ using standard normal multipliers. The empirical means and variances of ξ and ζ showed good compliance compared to the theoretical limit quantities stated in Section 3 (cf. Web Appendix S5).

Simultaneous 95% log-log transformed EP confidence bands for the probabilities of being alive and either under or free of IST within each treatment group are displayed in Figure 2 on the time interval $[t_1, t_2] = [4, 48]$ months. For comparison, log-log transformed 95% pointwise confidence

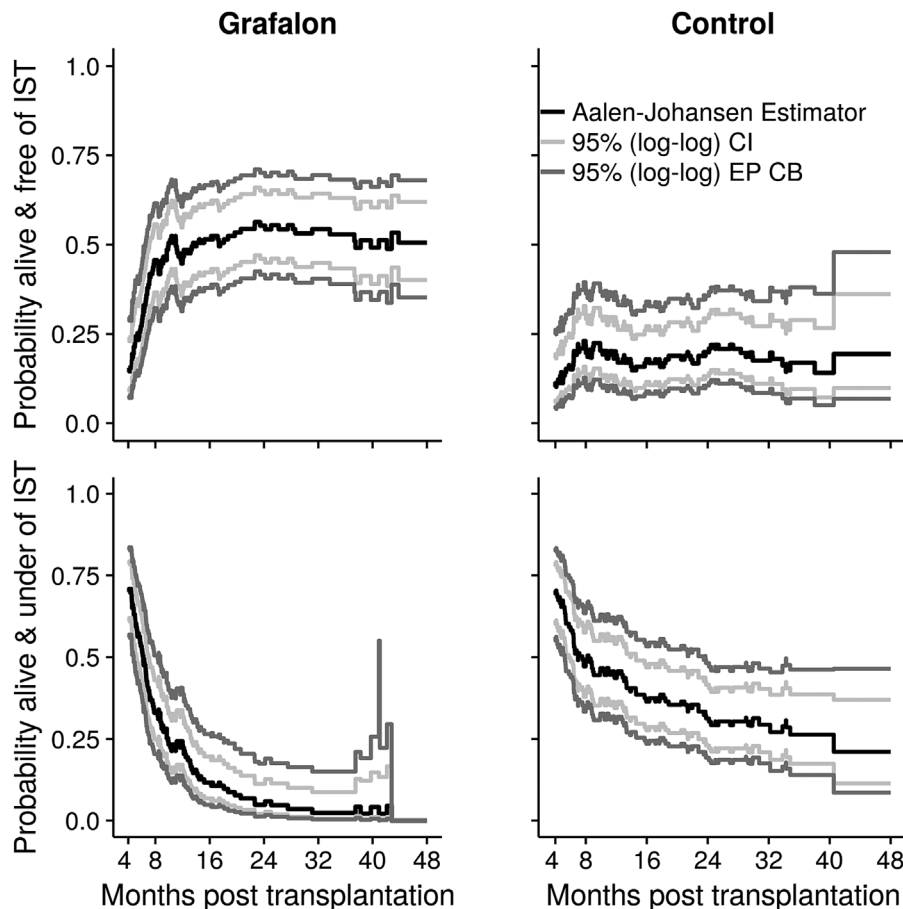


Figure 2. Upper panels: estimated probability to be alive and free of IST. Lower panels: estimated probability to be alive and under IST (black) for the Grafalon (left panels) and control group (right panels). 95% log-log pointwise confidence intervals (light gray) and log-log EP confidence bands (dark gray) in the time interval $[4, 48]$ months after transplantation are included.

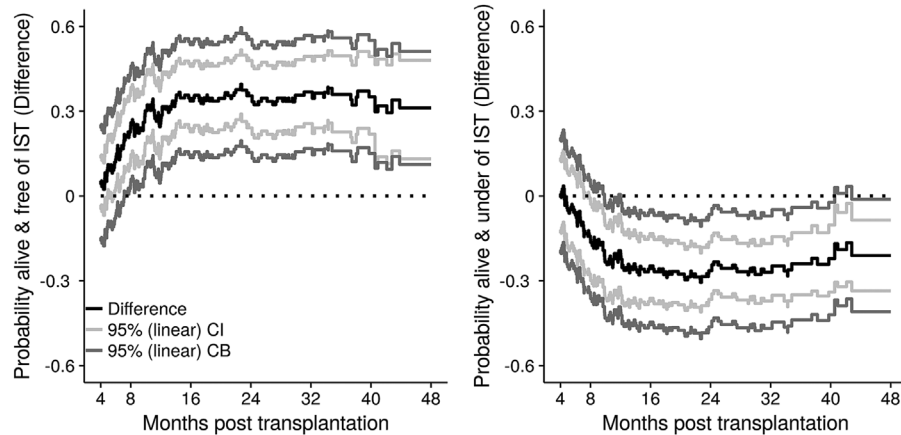


Figure 3. Left panel: estimated difference between Grafalon and control group regarding the probability to be alive and free of IST. Right panel: estimated difference between Grafalon and control group regarding the probability to be alive and under IST. 95% linear confidence bands (light gray) and 95% pointwise confidence intervals (dark gray) in the time interval $[4, 48]$ months after transplantation are included.

intervals are included. Confidence bands are only slightly broader than the pointwise 95% confidence intervals. The distinct peak of the upper confidence limit corresponding to the probability of being alive and under IST in Grafalon treated patients (lower left panel in Figure 2) is caused by instable weights for the latest timepoints due to small risk sets at the end of the study. Comparing the two treatment groups, the bands regarding both survival probabilities do not overlap starting at around 12 months post transplantation.

For statistical verification, Figure 3 displays their difference accompanied by the 95% linear confidence bands and 95% pointwise confidence intervals. We find a significant difference between Grafalon-treated patients and control patients, because both confidence bands exclude the “zero-effect” (horizontal dashed lines) for major parts of $t \in [4, 48]$. The bands demonstrate an increased probability of being alive and free of IST and a reduced probability of being alive and under IST as compared to control. The medical implication is that the addition of Grafalon to standard GvHD prophylaxis results in an increased proportion of patients being alive and not under IST after transplantation.

7. Discussion

The present article developed a wild bootstrap resampling technique for the Aalen–Johansen estimator for general time-inhomogeneous Markov multistate models. It allows for independent left-truncation, right-censoring as well as for degenerated initial distributions. The proposed approach first approximates the limiting distribution of the standardized Nelson–Aalen estimator. Afterward, the resulting quantities are transformed to approximate the limiting distribution of the Aalen–Johansen estimator. Compared to, for example, Lin (1997) and Beyersmann et al. (2013), the procedure considerably simplifies both computations and mathematical derivations, because involved martingale representations are avoided in favor of the much simpler representation of the Nelson–Aalen estimator. Contrary to the standard bootstrap approach with replacement, our technique does not require a strict iid setup, but allows for conditionally

independent trajectories given the initial state and possibly state-dependent left-truncation and right-censoring. Simulation results found satisfactory performance of the confidence bands in various settings. The applied log-log transformation improved coverage particularly for small sample sizes (results not shown). We investigated the performance of standard normal multipliers, which are the typical choice in biometrical applications, and centered Poisson multipliers with variance one. Differences between the two choices were negligible, and the choice of standard normal multipliers appears to be well-justified. We also found that 1000 bootstrap iterations are sufficient in the present scenarios. Adapting arguments given in Andersen et al. (1993) Section VII.2.3, covariates may be included by means of appropriate wild bootstrap resampling on the hazard scale as in a Cox model and subsequent transformation according to the Hadamard derivative. This is in contrast to, for instance, the resampling technique introduced for competing risks by Cheng et al. (1998). Further, our technique may also be used to derive other statistical tests than Kolmogorov–Smirnov-type tests, for instance, by adapting arguments given in Dobler and Pauly (2014). Scheike and Zhang (2003) also mentioned the present resampling idea as a possible way of inference in direct regression modeling. These are topics for further research.

The wild bootstrap enables a formal statistical framework for comparing complex time-to-event outcome probabilities, which are generally non-monotonic curves in time. In particular, the approach statistically confirms a positive treatment effect of Grafalon in terms of the time under immunosuppressive therapy under a nonparametric Markov assumption, but possibly requiring a slightly larger sample size. However, a sensitivity analysis with stricter nominal level $\alpha = 0.025$ supports our findings (results not shown). The proposed methodology has great potential in other fields of medical research, whenever statistical inference for transition probabilities or functionals thereof is required. Current examples include the clinical course of liver diseases (Jepsen et al., 2015), joint replacements in orthopaedic patients (Gillam et al., 2012), different stages of illicit drug use

(Mayet et al., 2012), pregnancy outcome data (Di Termini et al., 2012), investigations of longitudinal individual disability and mortality (Willekens, 2014), and infection control trials in healthcare epidemiology (e.g., Munoz-Price et al., 2016; Sommer et al., 2018, and references therein).

8. Supplementary Materials

Web Appendices, Tables, and Figures as well as Example R code referenced in Sections 3, 4, 5, and 6 are available with this article at the *Biometrics* website on Wiley Online Library.

ACKNOWLEDGEMENTS

The authors thank the referees and the editors for helpful comments that have substantially improved the article. The research leading to these results was conducted as part of the COMBACTE–MAGNET consortium. This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115523 | 115620 | 115737 resources of which are composed of financial contribution from the European Union Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution. Jan Beyersmann was partially supported by Grant BE 4500/1-1 of the German Research Foundation (DFG). Dennis Dobler and Markus Pauly were partially supported by the Strategic Research Fund (SFF) grant F-2012/375-12.

REFERENCES

- Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* **5**, 141–150.
- Allignol, A., Schumacher, M., and Beyersmann, J. (2011). Empirical transition matrix of multi-state models: The etm package. *Journal of Statistical Software* **38**, 1–15.
- Allignol, A., Schumacher, M., Wanner, C., Drechsler, C., and Beyersmann, J. (2011). Understanding competing risks: A simulation point of view. *BMC Medical Research Methodology* **11**, 1–13.
- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer Series in Statistics. New York, NY: Springer.
- Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing Risks and Multistate Models with R*. New York, USA: Springer Science & Business Media.
- Beyersmann, J., Termini, S. D., and Pauly, M. (2013). Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics* **40**, 387–402.
- Cheng, S., Fine, J. P., and Wei, L. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* **51**, 219–228.
- Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics* **146**, 162–169.
- Di Termini, S., Hieke, S., Schumacher, M., and Beyersmann, J. (2012). Nonparametric inference for the cumulative incidence function of a competing risk, with an emphasis on confidence bands in the presence of left-truncation. *Biometrical Journal* **54**, 568–578.
- Dobler, D., Beyersmann, J., and Pauly, M. (2017). Non-strange weird resampling for complex survival data. *Biometrika* **104**, 699–711.
- Dobler, D. and Pauly, M. (2014). Bootstrapping Aalen–Johansen processes for competing risks: Handicaps, solutions, and limitations. *Electronic Journal of Statistics* **8**, 2779–2803.
- Eefting, M., de Wreede, L. C., Halkes, C. J., Peter, A., Kersting, S., Marijt, E. W., et al. (2016). Multi-state analysis illustrates treatment success after stem cell transplantation for acute myeloid leukemia followed by donor lymphocyte infusion. *Haematologica* **101**, 506–514.
- Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312–319.
- Fiocco, M., Putter, H., and van Houwelingen, H. C. (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine* **27**, 4340–4358.
- Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics* **18**, 1501–1555.
- Gillam, M. H., Ryan, P., Salter, A., and Graves, S. E. (2012). Multi-state models and arthroplasty histories after unilateral total hip arthroplasties. *Acta Orthopaedica* **83**, 220–226.
- Jepsen, P., Vilstrup, H., and Andersen, P. K. (2015). The clinical course of cirrhosis: The importance of multistate models and competing risks analysis. *Hepatology* **62**, 292–302.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, New Jersey, USA: John Wiley & Sons.
- Lazaryan, A., Weisdorf, D. J., DeFor, T., Brunstein, C. G., MacMillan, M. L., Bejanyan, N., et al. (2016). Risk factors for acute and chronic graft-versus-host disease after allogeneic hematopoietic cell transplantation with umbilical cord blood and matched sibling donors. *Biology of Blood and Marrow Transplantation* **22**, 134–140.
- Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine* **16**, 901–910.
- Lin, D. Y., Fleming, T. R., and Wei, L.-J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.
- Lin, D. Y., Wei, L.-J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.
- Liu, L., Logan, B., and Klein, J. P. (2008). Inference for current leukemia free survival. *Lifetime Data Analysis* **14**, 432–446.
- Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics* **16**, 1696–1708.
- Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. New York, NY: Springer.
- Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. New York, NY: Springer.
- Mayet, A., Legleye, S., Falissard, B., and Chau, N. (2012). Cannabis use stages as predictors of subsequent initiation with other illicit drugs among french adolescents: Use of a multi-state model. *Addictive Behaviors* **37**, 160–166.
- Munoz-Price, L. S., Frencken, J. F., Tarima, S., and Bonten, M. (2016). Handling time-dependent variables: Antibiotics and antibiotic resistance. *Clinical Infectious Diseases* **62**, 1558–1563.
- Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics* **5**, 41–52.
- Pfirrmann, M., Hochhaus, A., Lauseker, M., Sauele, S., Hehlmann, R., and Hasford, J. (2011). Recommendations to meet statistical challenges arising from endpoints beyond overall survival in clinical trials on chronic myeloid leukemia. *Leukemia* **25**, 1433–1438.

- Scheike, T. H. and Zhang, M.-J. (2003). Extensions and applications of the Cox-Aalen survival model. *Biometrics* **59**, 1036–1045.
- Schmoor, C., Schumacher, M., Finke, J., and Beyersmann, J. (2013). Competing risks and multistate models. *Clinical Cancer Research* **19**, 12–21.
- Socié, G., Schmoor, C., Bethge, W., Ottinger, H. D., Stelljes, M., Zander, A. R., et al. (2011). Chronic graft-versus-host disease: Long-term results from a randomized trial on graft-versus-host disease prophylaxis with or without anti-T-cell globulin ATG-Fresenius. *Blood* **117**, 6375–6382.
- Sommer, H., Bluhmki, T., Beyersmann, J., and Schumacher, M. (2018). Assessing non-inferiority in treatment trials regarding severe infectious diseases: An extension to the entire follow-up period using a cure-death multistate model. *Antimicrobial Agents and Chemotherapy* **62**, e01691-17.
- Willekens, F. (2014). *Multistate Analysis of Life Histories with R*. New York, NY, USA: Springer International Publishing.

Received February 2017. Revised December 2017.

Accepted December 2017.