The Annals of Applied Statistics
2017, Vol. 11, No. 2, 840–867
DOI: 10.1214/17-AOAS1020

NONPARAMETRIC ESTIMATION OF PREGNANCY OUTCOME PROBABILITIES[1]

BY SARAH FRIEDRICH*, JAN BEYERSMANN*, URSULA WINTERFELD†, MARTIN SCHUMACHER‡ AND ARTHUR ALLIGNOL*

Ulm University, University Hospital Lausanne† and University Medical Center Freiburg‡*

Estimating pregnancy outcome probabilities based on observational cohorts has to account for both left-truncation, because the time scale is gestational age, and for competing risks, because, for example, an induced abortion may be precluded by a spontaneous abortion. The applied aim of this work was to investigate the impact of statins on pregnancy outcomes using data from Teratology Information Services. Using the standard Aalen–Johansen estimator of the cumulative event probabilities suggested the medically unexpected finding that statin exposure decreased the probability of induced abortion and led to more live births. The reason was an early induced abortion in a very small risk set in the control group, leading to unstable estimation which propagated over the whole time span. We suggest a stabilized Aalen–Johansen estimator which discards contributions from overly small risk sets. The decision whether a risk set is considered overly small is controlled by tuning parameters which we choose using a cross-validated Brier Score. We show that the new estimator enjoys the same asymptotic properties as the original Aalen–Johansen estimator. Small sample properties are investigated in extensive simulations. We also discuss extensions to more general multistate models.

1. Introduction. Prenatal development is the most vulnerable phase in human life. Drug toxicity, but also insufficiently treated maternal conditions may result in life-long handicaps of the newborn. Lupattelli et al. (2014) reported that around 80% of pregnant women use at least one drug to treat an acute or chronic condition. Teratology Information Services (TIS) advise both pregnant women and policy makers on the risk of adverse drug reactions in pregnancy. The societal impact is relevant: in Germany, for example, an annual number of about 900,000 pregnancies is assumed resulting in approximately 130,000 spontaneous abortions, more than 100,000 induced abortions, and 660,000 live births [Willand (2011, 2014)]. The aim of TIS counseling is to both reduce the rate of induced abortions based on irrational overestimation of drug risks and to lead to better and safer medical treatment in case of maternal disease [Hancock et al. (2007)]. In this context, TIS

Received April 2016; revised October 2016.

[1]This work was partially supported by Grant BE 4500/3-1 of the German Research Foundation (DFG).

Key words and phrases. Left-truncation, pregnancy, competing risks, abortion, Brier score.

observational cohort data are of outstanding value compared to register and prescription data which are often incomplete and not always reliable [Grzeskowiak, Gilbert and Morrison (2012)]. In particular, data on spontaneous abortion are not available from the latter sources and prescription is only a possibly biased proxy for actual treatment. Typically, pregnant women or their doctors contact TIS once the pregnancy has been recognized and drug risk assessment is needed. If consent is provided, and in addition to the individual counseling, TIS will then prospectively follow up pregnancies.

The statin study [Winterfeld et al. (2013)] enrolled pregnant women who—or whose physicians—contacted a TIS seeking advice on statin exposure during the first trimester of pregnancy. Follow up was achieved through a telephone interview or a mailed questionnaire to the woman or her physician after the expected delivery date. The control group consisted of women seeking advice on drugs known to be nonteratogenic.

As Meister and Schaefer (2008) pointed out, the statistical analysis of pregnancy outcomes is often based on standard multinomial estimates, that is, the number of outcome events divided by the number of pregnant women under study. The issue is that the natural time scale is gestational age, but women enter the study at different times after conception. The data are therefore left-truncated, that is, subject to delayed study entry. One consequence is that pregnancies that end in a spontaneous abortion before (in that case: hypothetical) study entry will not enter the cohort. The multinomial estimates do not account for this sampling aspect and, as Allignol, Schumacher and Beyersmann (2010) showed, will typically underestimate cumulative probabilities of both spontaneous and induced abortion. The reason is that the estimation bias is essentially mediated via underestimation of the abortion hazards. These are underestimated because ignoring left-truncation will artificially inflate risk sets.

Survival analysis accounts for left-truncation in that risk sets do not only decrease because of observed events or right-censorings, but may also increase because of delayed study entries. In a recent overview on the epidemiological study of fecundity and of pregnancy outcomes, Slama et al. (2014) stressed the need to use survival methodology that accounts for left-truncation. Andersen et al. (2012) studied the effect of alcohol intake during pregnancy on fetal death using prospective cohort data and argued that a strength of their investigation was the use of survival methods.

As again Meister and Schaefer (2008) pointed out, estimating pregnancy outcome probabilities does not only need to account for left-truncation, but also for competing risks, that is, different event types at the end of pregnancy. We will be particularly interested in spontaneous and induced abortions. The standard nonparametric estimator of the cumulative outcome probability in the presence of both left-truncation and competing risks is the Aalen–Johansen estimator [Aalen and Johansen (1978)]. One minus the sum of all the Aalen–Johansen estimators for all outcome types equals the Kaplan–Meier estimator for pregnancy duration. In the

absence of left-truncation and right-censoring, the Aalen–Johansen would equal the standard multinomial estimates.

However, and in contrast to the standard survival situation with only right-censoring, delayed study entries may lead to early random time intervals with small risk sets. Observed outcome events during such intervals may lead to unstable estimates which propagate over the whole time span. This appears to be what happened in a study on pregnancies exposed to statins [Winterfeld et al. (2013)]. As a consequence, the standard Aalen–Johansen estimator suggested that almost 40% of pregnancies in the control group ended in an induced abortion. This point estimate was not considered to be plausible, given, for example, the numbers cited for Germany above. As a further consequence, the point estimate of 40% induced abortions also implied a decreased risk of induced abortion and an increased proportion of live birth in the group of statin users, which was medically unexpected.

A common ad hoc approach—used by Winterfeld et al. (2013)—is to compute a conditional Aalen–Johansen estimator, conditional on no event until risk sets are large enough. However, this approach has drawbacks: First, it is ad hoc (unless the time point of conditioning has been specified in advance) and changes the interpretation. Second, one may possibly be faced with several disjoint time intervals of overly small risk sets.

The methodological aim of this work is to develop a stabilized Aalen–Johansen estimator, which is not ad hoc, and does not change the target quantity and accounts for problematic time intervals with overly small risk sets. The idea is to discard contributions from overly small risk sets, and traces back to Lai and Ying (1991) who modified the Kaplan–Meier estimator. We revisit their approach and show that its core is a modified Nelson–Aalen estimator [Andersen et al. (1993)]. Working with the fundamental Nelson–Aalen estimator allows for a generalization to competing risks. Our uniform consistency result complements that of Lai and Ying in that we do allow for a random number of random time intervals with overly small risk sets, while the proof of Lai and Ying essentially assumes that there is only one such early interval. Our weak convergence result allows for the same formalization of when a risk set is *overly small*. This is in contrast to Lai and Ying who imposed more restrictive assumptions for their weak convergence result. We also note that we use martingale arguments, complemented by the functional delta method, as in Andersen et al. (1993) throughout, while Lai and Ying used empirical process arguments for showing consistency. We also use martingale arguments to derive (co-) variance estimators, not derived in the earlier paper. To the best of our knowledge, our reanalysis of the statin data is one of the first real data analyses that illustrates the practical relevance of the Lai and Ying approach, now generalized to competing outcomes.

An important practical difficulty when using the suggestions of Lai and Ying and the present paper is that the concept of an overly small risk set is formalized as a function of the number of individuals under study and controlled by tuning parameters. For estimating pregnancy outcomes, we use a cross-validated choice

of the tuning parameters based on the Brier Score [e.g., Held and Sabanés Bové (2014)]. Predictions of pregnancy outcome are derived from the stabilized Aalen–Johansen estimates. The Brier Score is then estimated for updated predictions given study entry. In order to both train and apply the procedure on the same data set, this approach is combined with the 632 bootstrap [Gerds and Schumacher (2007)]. The paper is organized as follows: Section 2 introduces the competing risks model and the standard nonparametric Nelson–Aalen and Aalen–Johansen estimators. The modified estimators and their large sample properties are in Section 3. Extensive simulation results are reported in Section 4. The analyses of the statin data are in Section 5 and a discussion is in the final Section 6. The Appendix contains all proofs and considers (co-) variance estimation. Some discussion of the classical survival case and further simulation results are included in the supplement [Friedrich et al. (2017)].

2. The competing risks model. In the following, we will introduce the basic concepts and notations used in this work.

In a competing risks setting, we consider $k \geq 2$ competing absorbing states. For ease of presentation, we will assume two competing states in the following.

The stochastic process $(X(t), t \geq 0)$ corresponding to Figure 1 gives the state occupied by the individual at time t, that is, $X(t) = 0$ as long as no event has happened and $X(t) = j$, if an event of type j, $j = 1, 2$, has occurred in $[0, t]$. The event time is defined as

$$T = \inf\{t : X(t) \neq 0\},$$

that is, T denotes the smallest time at which the process is not in the initial state anymore.

The type of the first event, often called cause of failure, is given by $X(T) \in \{1, 2\}$, that is, the state the process enters at time T.

Key quantities are the cause-specific hazards $\alpha_{0j}(t)$ for each competing event j,

$$\alpha_{0j}(t) = \lim_{\Delta t \searrow 0} \frac{P(T \in [t, t + \Delta t), X(T) = j | T \geq t)}{\Delta t}, \tag{2.1}$$

which we assume to exist. Furthermore, we presume that $\int_0^\tau \alpha_{0j}(u)\,du < \infty$ for some time interval $[0, \tau]$, $\tau < \infty$. The cumulative cause-specific hazards are

$$A_{0j}(t) = \int_0^t \alpha_{0j}(u)\,du, \qquad j = 1, 2.$$

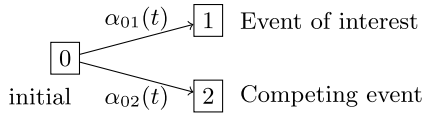

FIG. 1. *Competing risks model with two competing events and cause-specific hazards* α_{0j}, $j = 1, 2$.

The cause-specific hazards can be thought of as momentary forces of transition, moving along the arrows in Figure 1. It is crucial to note that both cause-specific hazards completely determine the behavior of the competing risks process [Andersen et al. (1993), Chapter II.6].

Another quantity of interest is the cumulative incidence function (CIF), which describes the transition probability from state 0 to state j by time t:

$$P_{0j}(0,t) = P\big(T \le t, X(T) = j\big)$$

for $j = 1, 2$. More general, the matrix of transition probabilities is given by

$$\mathbf{P}(s,t) = \big(P_{lm}(s,t)\big)_{l,m},$$

where $P_{lm}(s,t) = P(X(t) = m | X(s) = l)$, $l, m \in \{0, 1, 2\}$. The transition probability matrix may be written as a functional of the cumulative hazards via product integration [Andersen et al. (1993), Theorem II.6.7]:

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

where $\mathbf{A} = (A_{lm})_{l,m}$, $A_{ll}(t) = -\sum_{m=0, l\neq m}^{2} A_{lm}(t)$ and $\mathbf{I}$ is the identity matrix.

2.1. *Nonparametric estimation.* Usually, an individual's event time and cause of failure will be subject to right-censoring and/or left-truncation. TIS observational cohort data are typically only left-truncated such that pregnancy outcomes are observed for every woman under study. Hence, our main focus will be on left-truncation, similar to Keiding and Gill (1990). For completeness, however, we have included right-censoring in our theoretical derivations.

Left-truncation arises, if study entry happens at some time point later than time origin. For the case of pregnancy outcomes time origin would be the last menstrual period, but women enter the study several weeks after conception. If the pregnancy ends before a woman enters the study, it will never be observed.

In the presence of both left-truncation times L_i and right-censoring times C_i, the observable data for individual i consists of $(\tilde{T}_i, \delta_i, X(\tilde{T}_i))$. Here, $\tilde{T}_i = (T_i \wedge C_i)$ denotes the minimum of event and censoring time with $L_i < \tilde{T}_i$ and $\delta_i = \mathbb{1}(L_i < T_i \le C_i)$. Recall that for TIS observational cohort data we will typically have $C_i = \infty$ for all individuals i under study. The at-risk process and counting processes may be formulated as follows. For individual i with left-truncation time L_i and observed event time $\tilde{T}_i$ as well as event type j, we have

$$N_{0j;i}(t) = \mathbb{1}\big(L_i < \tilde{T}_i \le t, X(\tilde{T}_i) = j, \delta_i = 1\big),$$

$$Y_i(t) = \mathbb{1}(L_i < t \le \tilde{T}_i).$$

Here, $N_{0j;i}(t)$ counts the number of observed $0 \to j$ transitions for individual i in $[0, t]$. Aggregating over all n individuals under study gives

$$N_{0j}^{(n)}(t) = \sum_{i=1}^{n} N_{0j;i}(t),$$

as well as the at-risk process

$$Y^{(n)}(t) = \sum_{i=1}^{n} \mathbb{1}(L_i < t \le \tilde{T}_i).$$

For ease of presentation, we will drop the superscript n in the following.

The Nelson–Aalen estimators of the cumulative cause-specific hazards [Andersen et al. (1993), Section IV.1] are given by

$$\tilde{A}_{0j}(t) = \sum_{s \le t} \frac{\Delta N_{0j}(s)}{Y(s)}$$

for $j = 1, 2$ and the sum is over all observed, unique event times $s \le t$.

The cumulative incidence functions may be estimated by the Aalen–Johansen estimator:

$$\tilde{P}_{0j}(0, t) = \sum_{s \le t} \tilde{S}(s-) \frac{\Delta N_{0j}(s)}{Y(s)}, \tag{2.2}$$

where

$$\tilde{S}(t) = \prod_{s \le t} \big(1 - \Delta \tilde{A}_{0\cdot}(s)\big) \tag{2.3}$$

is the Kaplan–Meier estimator, that is, the estimated probability of the waiting time T in the initial state 0 to exceed time t. The product is taken over all observed, unique event times $s \le t$.

The connection of the general competing risks framework to the statin study is presented in detail in the supplemental material [Friedrich et al. (2017)].

3. Modified Nelson–Aalen and Aalen–Johansen estimators for competing risks. We will now present a modification of the Nelson–Aalen estimator to avoid the problems caused by small risk sets at the beginning of a study, but possibly also during intermediate time intervals.

Since for left-truncated data the number of individuals is not known at time 0, we consider the data as being generated by a larger sample $\tilde{T}_i, L_i, i = 1, \ldots, m(n)$, where

$$m(n) = \inf\left\{ m : \sum_{i=1}^{m} \mathbb{1}(L_i < \tilde{T}_i) = n \right\} \tag{3.1}$$

as in Lai and Ying (1991). Here, n is the number of individuals under study.

To avoid the problems caused by small risk sets, we define the modified Nelson–Aalen estimator as $\hat{\mathbf{A}} = (\hat{A}_{01}, \hat{A}_{02})$, where

$$\hat{A}_{0j}(t) = \sum_{s \le t} \frac{\Delta N_{0j}(s)}{Y(s)} \mathbb{1}\big(Y(s) \ge cn^{\gamma}\big) \tag{3.2}$$

for $j = 1, 2$, $c > 0$ and $\gamma \in (0, 1)$. In the absence of competing risks, the modification in (3.2) leads to the Kaplan–Meier-type estimator of Lai and Ying (1991) using product integration as in Section 3.2 below. Our approach is based on $\hat{\mathbf{A}}$ first, because the Kaplan–Meier estimator must not be used for estimating CIFs [Gooley (1999)].

3.1. *Large sample properties of the modified Nelson–Aalen estimator.* In this section, we will establish uniform strong consistency of the modified Nelson–Aalen estimator from (3.2) as well as weak convergence toward a Gaussian martingale on the space of càdlàg functions $D[0, \tau]^2$ using martingale arguments.

Since the only difference between the Nelson–Aalen estimator and its modification is the indicator function $\mathbb{1}(Y(s) \geq cn^{\gamma})$, the argumentation seems to be straightforward at first sight. For martingale arguments to work, however, we need $\mathbb{1}(Y(s) \geq cn^{\gamma})$ to be predictable. But since n is random and not known at the beginning of the study, $\mathbb{1}(Y(s) \geq cn^{\gamma})$ is not predictable with respect to the usual filtration generated by the past observed data. In order to avoid this problem, we need to consider a different filtration—proposed by Lai and Ying (1991) for their weak convergence result—which will be introduced in the following.

Let $\mathcal{F}(s)$ be the σ-field generated by

$$L_i, \mathbb{1}(L_i < \tilde{T}_i), \mathbb{1}(L_i < u \leq \tilde{T}_i), \mathbb{1}(L_i < \tilde{T}_i \leq u), \delta_i X(\tilde{T}_i)\mathbb{1}(L_i < \tilde{T}_i \leq u) \tag{3.3}$$

for $u \leq s$ and $i = 1, 2, \ldots, m(n)$. This means that in addition to the observed past, left-truncation times and whether or not an individual enters the study are assumed to be known for all individuals and all time points. Additionally, the at-risk status and the vital status are known for individuals under study and for all times $u >$ study entry.

The awkward aspect of this construction is that by assuming all left-truncation times and all study entry statuses to be known beforehand, we possibly violate the principle that one shall not condition on the future in an event history analysis [Andersen and Keiding (2012)]. However, the only practical difference between using this new filtration $\mathcal{F}(s)$ instead of the usual one is the fact that, by conditioning on the study entry statuses $\mathbb{1}(L_i < \tilde{T}_i)$ of all individuals, n is not random anymore. This implies that $m(n)$ and $\mathbb{1}(Y(s) \geq cn^{\gamma})$ are both predictable with respect to $\mathcal{F}(s-)$ [Lai and Ying (1991), Lemma 5].

ASSUMPTION 3.1. In the following, we will assume that the intensities of the counting processes have a multiplicative intensity structure w.r.t. $\mathcal{F}(t)$ as in Andersen et al. (1993), Section IV.1.2, such that

$$M_{0j}(t) = N_{0j}(t) - \int_0^t Y(s)\alpha_{0j}(s)\,ds, \qquad j = 1, 2,$$

are martingales w.r.t. $\mathcal{F}(t)$, using predictability of $\mathbb{1}(Y(s) \geq cn^{\gamma})$.

Practically speaking, we assume that the knowledge of whether or not more patients will enter the study later does not change the intensity of having an event for a single individual. In particular, this assumption is fulfilled, if we assume i.i.d. competing risks data, observation of which is subject to random left-truncation and right-censoring, which is reasonable for many practical applications. Note, however, that this assumption is stricter than conditioning on the usual filtration; see Remark 7.1 in the supplemental material for an example.

The templates for our theorems and proofs are Theorems IV.1.1–2 of Andersen et al. (1993), but using $\mathcal{F}(t)$ because of the modifications in (3.2).

We can now formulate and prove uniform strong consistency and weak convergence for the modified Nelson–Aalen estimator:

THEOREM 3.2 (Strong consistency). *Let $t \in [0, \tau]$, $c > 0$ and $\gamma \in (0, 1)$ and assume that, as $n \to \infty$,*

$$\int_0^t \frac{\mathbb{1}(Y(s) \geq cn^\gamma)}{Y(s)} \alpha_{0j}(s)\, ds \xrightarrow{P} 0 \tag{3.4}$$

and

$$\int_0^t \big(1 - \mathbb{1}\big(Y(s) \geq cn^\gamma\big)\big)\alpha_{0j}(s)\, ds \xrightarrow{P} 0. \tag{3.5}$$

Then, as $n \to \infty$,

$$\sup_{s \in [0,t]} \big|\hat{A}_{0j}(s) - A_{0j}(s)\big| \xrightarrow{P} 0$$

for $j = 1, 2$.

This implies that

$$\|\hat{\mathbf{A}} - \mathbf{A}\| \xrightarrow{P} 0 \qquad \text{as } n \to \infty,$$

where $\|\cdot\|$ is the max-supremum norm.

PROOF. The proof is deferred to the Appendix. □

THEOREM 3.3 (Weak convergence). *Let $t \in [0, \tau]$, $c > 0$, $\gamma \in (0, 1)$ and assume that there exist nonnegative functions $y(s)$ such that $\alpha_{0j}(s)/y(s)$ is integrable on $[0, t]$ for $j = 1, 2$. Let*

$$\sigma_j^2(t) = \int_0^t \frac{\alpha_{0j}(s)}{y(s)}\, ds \tag{3.6}$$

and assume that:

(1) *For each $s \in [0, t]$ and $j = 1, 2$,*

$$n \int_0^s \frac{\mathbb{1}(Y(u) \geq cn^\gamma)}{Y(u)} \alpha_{0j}(u)\, du \xrightarrow{P} \sigma_j^2(s) \qquad \text{as } n \to \infty.$$

(2) *For* $j = 1, 2$ *and all* $\varepsilon > 0$,

$$n \int_0^t \frac{\mathbb{1}(Y(u) \geq cn^\gamma)}{Y(u)} \alpha_{0j}(u) \mathbb{1}\left(\left| \sqrt{n} \frac{\mathbb{1}(Y(u) \geq cn^\gamma)}{Y(u)} \right| > \varepsilon \right) du \xrightarrow{P} 0$$

as $n \to \infty$.

(3) *For* $j = 1, 2$,

$$\sqrt{n} \int_0^t \big(1 - \mathbb{1}\big(Y(u) \geq cn^\gamma\big)\big) \alpha_{0j}(u)\, du \xrightarrow{P} 0 \qquad \text{as } n \to \infty.$$

Then

$$\sqrt{n}(\hat{\mathbf{A}} - \mathbf{A}) \xrightarrow{\mathcal{D}} \mathbf{W} = (W_1, W_2)$$

in $D[0, \tau]^2$ *as* $n \to \infty$, *where* W_1, W_2 *are independent Gaussian martingales with covariance function*:

$$\operatorname{Cov}\big(W_j(s_1), W_j(s_2)\big) = \sigma_j^2(s_1 \wedge s_2).$$

Here, $s_1 \wedge s_2$ *denotes the minimum of* s_1 *and* s_2.

PROOF. The proof is in the Appendix. □

Note that the conditions (3.4) and (3.5) are fulfilled, if $Y(s)/n$ is uniformly bounded away from 0 on $[0, \tau]$ in probability, w.r.t. a probability measure given study entry; see Example IV.1.7 in Andersen et al. (1993). Sufficient conditions for this are i.i.d. event times T_i with absolutely continuous distribution function $F(t)$ such that $F(s) < 1$ for all $s \in [0, \tau]$ and i.i.d. left-truncation times and censoring times (L_i, C_i), independent of the T_i's. Let $G(s) = P(L < s \leq C)$ and assume that $G(s) > 0$ for all $s \in [0, \tau]$ and $L_i < C_i$ with probability 1. In this case, Conditions (1)–(3) of Theorem 3.3 are fulfilled and $y(s)$ is given by

$$y(s) = \frac{1}{p} G(s)\big(1 - F(s)\big),$$

where $p = P(L < T)$. Note that this covariance function of the limiting Gaussian martingale is the same as for the usual Nelson–Aalen estimator; see Andersen et al. (1993), Example IV.1.7.

3.2. *The modified Aalen–Johansen estimator for competing risks.* We can use the modified Nelson–Aalen estimator for competing risks from (3.2) to define a modification of the Aalen–Johansen estimator via product integration:

$$\hat{\mathbf{P}}(s, t) = \prod_{(0,t]} (\mathbf{I} + d\hat{\mathbf{A}}) = \prod_{s \leq t} \big(\mathbf{I} + \Delta \hat{\mathbf{A}}(s)\big),$$

where $\mathbf{I}$ is the 2×2 identity matrix and the product is over all observed, unique event times $s \leq t$. Strong consistency and weak convergence for the modified

Aalen–Johansen estimator then follow by using the continuous mapping theorem and the functional delta method, similar to Andersen et al. (1993), Section IV.4.2. Again, note that the limit distribution is the same as for the classical Aalen–Johansen estimator and will therefore not be repeated here.

For the estimated cumulative incidence functions, we get the following modification:

$$\hat{P}_{0j}(0,t) = \sum_{s \le t} \hat{S}(s-) \frac{\Delta N_{0j}(s)}{Y(s)} \mathbb{1}\big(Y(s) \ge cn^{\gamma}\big), \tag{3.7}$$

where $\hat{S}(t)$ is the modified Kaplan–Meier estimator for overall survival due to Lai and Ying (1991),

$$\hat{S}(t-) = \prod_{u<t} \left(1 - \frac{\Delta N_{0\cdot}(u)}{Y(u)} \mathbb{1}\big(Y(u) \ge cn^{\gamma}\big)\right);$$

see the supplemental material [Friedrich et al. (2017)] for a derivation of the modified Kaplan–Meier estimator in the classical survival case.

3.3. *Choice of the tuning parameters c and γ.* The modified Aalen–Johansen estimator is based on tuning parameters c and γ which—in a real data analysis—need to be chosen properly. We propose to use a cross-validation procedure [e.g., Hastie, Tibshirani and Friedman (2009)] combined with a 632 bootstrap approach [Gerds and Schumacher (2007)] that will be explained in Section 4.3. The aim of the present subsection is to suggest a measure of prediction error for pregnancy outcome data which can be used in the cross-validation process.

To this end, recall that survival methods are used because pregnancy cohorts are left-truncated, but primary interest is in the eventual outcome $X(T)$. We estimate $P(X(T) = j)$ by the right-hand limit $\hat{P}_{0j}(0, \infty)$ of the (modified) Aalen–Johansen estimator. The aim is now to compare the prediction of a type j event for all n individuals i in the data with the actual event status $X(T_i)$. Because individual i has left-truncation time $L_i = l_i$, we use the updated prediction [van Houwelingen and Putter (2012), pp. 45–46]

$$\pi_i = \hat{P}\big(T \le \infty, X(T) = j | T > l_i\big) = \frac{\hat{P}(T > l_i, T \le \infty, X(T) = j)}{\hat{S}(l_i)}, \tag{3.8}$$

$i = 1, \ldots, n$, and consider the estimated Brier Score [e.g., Held and Sabanés Bové (2014)]:

$$BS = \frac{1}{n} \sum_{i=1}^{n} \big(\mathbb{1}\big(X(T_i) = j\big) - \pi_i\big)^2 \tag{3.9}$$

as a measure of prediction error.

Estimating the Brier Score for survival data is typically complicated by right-censoring, because the event status after censoring is unknown. This has been

addressed by Inverse Probability of Censoring Weighting for time-to-death data [Gerds and Schumacher (2006)] and for competing risks [Schoop et al. (2011)]. In (3.9), we exploit that pregnancy outcomes are observed for all women under study, which in turn implies that (3.9) estimates the expected Brier Score w.r.t. a probability measure given study entry. This is common for left-truncated data [Andersen et al. (1993), Example IV.1.7].

4. Simulation studies. We will now conduct several simulation studies in order to analyze the behavior of our modified Aalen–Johansen estimator in different scenarios. We begin by describing the simulation setting, which is custom-made to produce events in early small risk sets, similar to the pregnancy data that will be analyzed in detail in Section 5. Next, and as suggested by Lai and Ying (1991), we focus on the choice $\gamma = 1/4$ and $c = 1$ for the tuning parameters and analyze the small sample behavior in this scenario. Finally, the tuning parameters are chosen using cross-validation combined with a resampling approach. In addition, the supplement [Friedrich et al. (2017)] presents results concerning large sample consistency as well as a second simulation setting with less pronounced truncation.

Results of 10,000 simulation runs are reported as plots of the estimated CIFs, comparing both the classical Aalen–Johansen estimator and its modified version to the real CIF. Furthermore, the average number of individuals at risk, the bias, relative bias and the root mean squared errors (RMSE) for both estimators as well as the variance estimators are reported in tabular form in Section 9 of the supplemental material.

4.1. *Simulation setting.* Our aim is to simulate competing risks data similar to the pregnancy data from Section 5. Since our interest lies in what happens at the beginning of the study, we only simulate two competing states that might be interpreted as spontaneous and induced abortion. Motivated by the data example from Section 5, we chose a linearly decreasing cause-specific hazard

$$\alpha_{01}(t) = -1.7 \cdot 10^{-4} \cdot t + 0.017$$

for the event of interest and a Weibull-type cause-specific hazard

$$\alpha_{02}(t) = \frac{1.4}{27^{1.4}} \cdot t^{0.4}$$

for the competing event. Competing risks data were simulated as in Beyersmann et al. (2009):

1. Survival times T are simulated with all-cause hazard $\alpha_{0\cdot}(t) = \alpha_{01}(t) + \alpha_{02}(t)$
2. Given a survival time $T = t$, a binomial experiment is run, which decides on cause 1 with probability

$$P\big(X(T) = 1 | T = t\big) = \frac{\alpha_{01}(t)}{\alpha_{0\cdot}(t)}$$

3. Left-truncation times L are simulated independently of $[T, X(T)]$.

The truncation times followed a skewed normal distribution with density function

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left[\frac{x-\xi}{\omega}\right]\right),$$

where ξ is the location parameter, ω the scale parameter and α the shape parameter and ϕ and Φ are the density and the cumulative distribution function of a standard normal distribution, respectively. The parameters were chosen as $(\xi, \omega, \alpha) = (16, 4.3, -8)$.

We simulated two different scenarios to analyze the behavior in small samples as well as in large samples (see the supplement [Friedrich et al. (2017)] for details and results of the latter).

4.2. *Small sample behavior.* For the small sample scenario, $m = 200$ individuals were simulated. Since the data are left-truncated, only individuals with $L < T$ enter the study. Therefore, the number n of people under study was random with $n \leq m$, ranging from 89 to 144. On average, $n = 117$ people entered the study in the small sample scenario.

The value of $\lceil cn^{\gamma} \rceil$ was calculated for each data set with the corresponding number of individuals under study in this data set. In the plots below, the average value of $\lceil cn^{\gamma} \rceil$ is reported, which is stressed by using, for example, $\lceil cn^{\gamma} \rceil \approx 4$ instead of $\lceil cn^{\gamma} \rceil = 4$. The same notation is used for n. For the simulated data sets, the Aalen–Johansen estimator for the event of interest and its modified version were calculated along with 95% complementary log-minus-log transformed confidence intervals (CI) based on the usual and the modified variance estimator, respectively. Both variance estimators were of the Greenwood-type and calculated according to Section A.1 in the Appendix. The complementary log-minus-log transformed confidence interval takes the form [Beyersmann, Allignol and Schumacher (2012)]

$$1 - \left(1 - \hat{P}_{0j}(0,t)\right)^{\exp\left(\pm z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_{\mathrm{AJ}}(t)}{(1-\hat{P}_{0j}(0,t))\cdot\log(1-\hat{P}_{0j}(0,t))}\right)}, \tag{4.1}$$

where $\hat{P}_{0j}(0,t)$ is the Aalen–Johansen estimator or its modified version, respectively, and $\hat{\sigma}^2_{\mathrm{AJ}}(t)$ is the corresponding estimated variance. The transformation guarantees that the CI is contained in $[0, 1]$.

For the plots below, however, we used empirical point-wise 95% confidence intervals constructed by taking the 2.5% and 97.5% percentile of the simulated CIFs. Note that we have inserted a small gap between the estimators in all plots to improve distinguishability, but the estimators were calculated for the same time points.

Figure 2 shows that both estimators underestimate the true CIF for the event of interest, but the median bias of the two estimators is comparable. However,

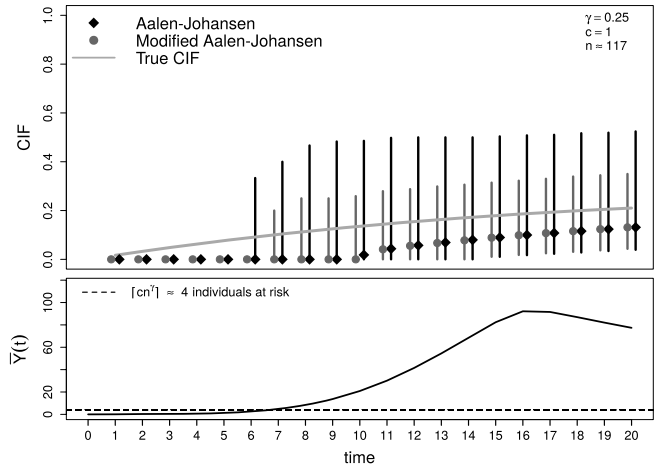

FIG. 2. *Results of* 10,000 *simulation runs in the small sample scenario*: *On average*, $n = 117$ *out of* 200 *simulated individuals enter the study. Displayed are the median Aalen–Johansen estimator and its modified version along with empirical* 95% *CIs*, *as well as the true CIF for the event of interest* (*gray line*). *The lower plot shows the average number of individuals at risk over the course of time.*

the empirical confidence intervals for the new estimator are smaller than the ones for the Aalen–Johansen estimator. An explanation can be found by looking at the mean bias; see Figure 7 in the supplement. Here, we see that the modified Aalen–Johansen estimator has—on average—a larger absolute bias than the usual Aalen–Johansen estimator. As we can see in Figure 8 in the supplement, this is due to the fact that the modified estimator eliminates "outliers." The light gray lines in Figure 8 display randomly chosen CIFs, estimated by the original Aalen–Johansen estimator. We see a large amount of these reaching implausibly high levels at the plateau, because an event happened early, adding to the variability of the usual Aalen–Johansen estimator. The modified estimator, in contrast, ignores these "outliers," resulting in a higher absolute bias *on average* and a smaller variation.

A detailed analysis for the scenario with $\gamma = 0.25$ and $c = 1$ is presented in Table 9.1 in the supplement. Reported are the average number of individuals at risk $\bar{Y}(t)$, the mean and median bias for both the true Aalen–Johansen estimator and the modified version, the root mean squared errors (RMSE) for both estimators as well as the variance estimators. The RMSE is computed as $\sqrt{(\bar{\hat{F}} - F)^2 + \operatorname{var}(\hat{F})}$, where F denotes the true CIF for the event of interest and $\bar{\hat{F}}$ is the averaged CIF estimated by either the Aalen–Johansen estimator or the modified estimator.

More extreme choices of c and γ lead to meaningless estimates; see Section 9.1 in the supplemental material. There, we also included the large sample simulations, which confirm consistency of the new estimator.

In order to analyze the behavior of the new variance estimator, we plotted the variance estimators for both the Aalen–Johansen and the new estimator as well

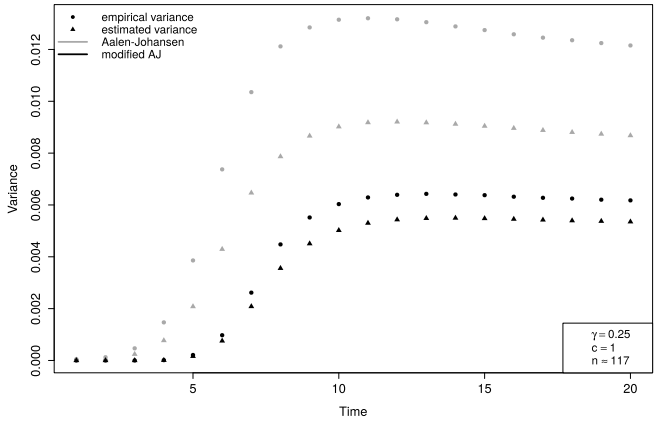

FIG. 3. *Results of* 10,000 *simulation runs in the small sample scenario*: *On average*, $n = 117$ *out of* 200 *simulated individuals enter the study. Displayed are the mean variance estimators of the Aalen–Johansen estimator and its modified version*, *respectively*, *compared to the corresponding empirical variance.*

as the corresponding empirical variances for $\gamma = 0.25$, $c = 1$ and $n = 117$. Both variance estimators are biased downward compared to the empirical variance; see Figure 3. Such negative bias has also been documented for the Kaplan–Meier estimator [Klein (1991)]. As anticipated, Figure 3 also documents the smaller variation of the new estimator.

4.3. *A cross-validated and resampled choice of c and* γ. So far, we considered ad hoc choices for the tuning parameters c and γ. In order to determine suitable—data-driven—choices for these parameters we have used cross-validation combined with a resampling approach in order to both train and apply the stabilized estimator on the same data set. This procedure will be described in detail in the following.

We restricted our investigation to the small sample simulation setting in the above scenario, that is, we simulated 10,000 data sets, each with $m = 200$ individuals out of which on average $n = 117$ entered the study. For each data set, we used a cross-validation procedure to choose the tuning parameters, say c_0 and γ_0, which were then used for estimation of the cumulative incidence function.

Since a classical cross-validation estimator tends to be positively biased because it is trained with less information than provided by the full data [Gerds and Schumacher (2007)], we have applied a 632 bootstrap procedure. The idea behind this approach is to balance the upward bias (resulting from using fewer data than in the full data set) and the downward bias of the so-called apparent error [Efron (1983), Efron and Tibshirani (1997), Gerds and Schumacher (2007)]. The procedure works as follows.

TABLE 1
Choices for c and γ such that $\lceil cn^{\gamma} \rceil \in \{1, \ldots, 10\}$ for $n = 117$

$\lceil cn^{\gamma} \rceil$	c	γ
1	0.02	0.75
2	0.7	0.1
3	2	0.01
4	1	0.25
5	0.3	0.55
6	0.5	0.5
7	1.5	0.3
8	0.1	0.9
9	6	0.07
10	3.5	0.2

For each simulated data set, we first drew $B = 1000$ bootstrap samples of size n with replacement from the data. For each of the $b = 1, \ldots, B$ samples Q_b^*, we calculated the modified Aalen–Johansen estimator $\hat{\mathbf{P}}$ with parameters (c, γ) for a choice of parameters as in Table 1.

Following Section 3.3, we then predicted the pregnancy outcome for each individual i *not* in Q_b^* by

$$\pi_i = \frac{\hat{P}(T > l_i, T \leq \infty, X(T) = 1)}{\hat{S}(T > l_i)},$$

where l_i denotes the left-truncation time of individual i and $j = 1$ is the event of interest.

The predicted outcomes were compared to the true outcomes by estimating the Brier Score for all $\tilde{n}$ individuals not in Q_b^*, that is,

$$\widehat{\mathrm{Err}}_b = \frac{1}{\tilde{n}} \sum_{i \notin Q_b^*} \big(\mathbb{1}\big(X(T_i) = 1\big) - \pi_i\big)^2.$$

We then averaged over all bootstrap samples, that is, we calculated

$$\widehat{\mathrm{Err}}_{B_0} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathrm{Err}}_b.$$

Finally, the resulting error is given as

$$\widehat{\mathrm{Err}}_{\omega} = (1 - \omega) \frac{1}{n} \sum_{i=1}^{n} \big(\mathbb{1}\big(X(T_i) = 1\big) - \pi_i\big)^2 + \omega \widehat{\mathrm{Err}}_{B_0},$$

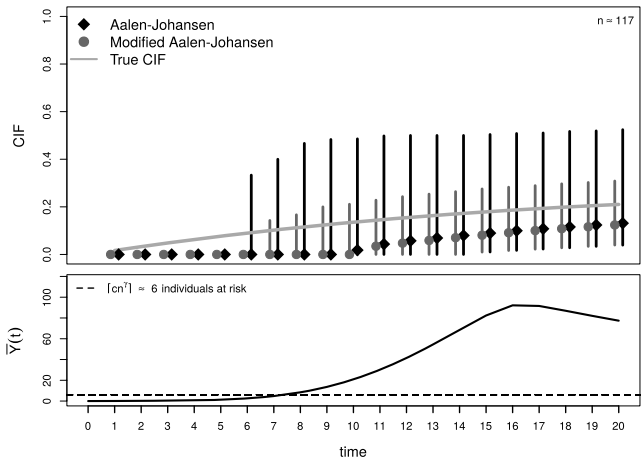

FIG. 4. *Results of* 10,000 *simulation runs in the cross-validation study*: *Displayed are the median Aalen–Johansen estimator and its modification using* (c_0, γ_0) *along with* 95% *empirical CIs and the number of individuals at risk over the course of time* (*lower plot*). *On average*, $n = 117$ *out of* 200 *simulated individuals enter the study.*

with $\omega = 0.632$, which balances the upward bias of $\widehat{\mathrm{Err}}_{B_0}$ and the downward bias of the apparent error $\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}(X(T_i) = 1) - \pi_i)^2$ [see Efron (1983), Efron and Tibshirani (1997), Gerds and Schumacher (2007) for further details].

These steps were conducted for all parameters (c, γ) displayed in Table 1. Here, we chose pairs of parameters such that $\lceil cn^{\gamma} \rceil$ covered the values between 1 and 10 for $n = 117$, that is, the border for an overly small risk set varied from 1 to 10. The parameters (c_0, γ_0) resulting in the smallest $\widehat{\mathrm{Err}}_{\omega}$ were used for the calculation of the results. Combining all 10,000 simulation runs resulted in a median value of $\lceil cn^{\gamma} \rceil \approx 6$. Figure 4 illustrates findings similar to Section 4.2 for the original choice of $\gamma = 0.25$ and $c = 1$. A plot of the mean estimators is in Section 9.3 of the supplement.

4.4. *Coverage probabilities*. Additionally, we calculated the coverage probabilities for the small sample scenario. Therefore, the complementary log-minus-log transformed confidence intervals were calculated for the classical Aalen–Johansen estimator as well as for the modified estimator, once with $\gamma = 0.25$ and $c = 1$ and once with the cross-validated (c_0, γ_0). The percentage of simulation studies, in which the estimated CI contained the true value of the CIF is reported in Table 2. As one can see, the coverage probabilities for all estimators are comparable, but smaller than 95%.

The implication is threefold: As seen above, the new estimator protects against the CIs being constructed around implausibly large values. The width of the empirical CIs is smaller compared to the standard method, while coverage probabilities are similar and do not reach the nominal level.

TABLE 2

Coverage probabilities (CP) from 10,000 *simulation studies in % for the modified and the original Aalen–Johansen estimator, respectively. Complementary log-minus-log transformed* 95% *CIs used, mean* $n = 117$. *CP modified*: $\gamma = 0.25$, $c = 1$, *CP cross-validation*: $\lceil cn^{\gamma} \rceil \approx 6$. *Right column*: *CP of the modified estimator in a scenario with less pronounced left-truncation*, $n \approx 184$, $\gamma = 0.25$, $c = 1$; *see the supplement for details*

	Small sample scenario			Light truncation scenario
Time	**CP modified**	**CP cross-validation**	**CP original**	**CP modified**
2	0.00	0.00	0.00	23.21
4	0.02	0.02	0.11	82.17
6	2.03	0.70	3.62	92.25
8	16.66	9.75	20.25	91.43
10	46.95	41.62	49.14	90.88
12	71.19	68.69	71.31	90.92
14	68.92	66.27	69.38	91.35
16	67.03	64.51	67.97	91.73
18	67.28	64.92	68.39	91.85
20	68.29	66.15	69.56	92.02

For our application, the main interest lies in the behavior at the plateau, that is, coverage probabilities of only 70% at time 20 are a concern whereas the small coverage probabilities for early time points ($< 5\%$ for $t = 6$) are not.

For comparison, coverage probabilities of the modified estimator in a scenario with less pronounced truncation are displayed in the right column of Table 2; see Section 9.4 in the supplement for details on the simulation setting. As we can see, coverage probabilities are much closer to the nominal level in this scenario. Furthermore, coverage probabilities were also calculated in a simulation setting without truncation. We found that the nominal level of 95% was always approximately reached (results not shown). This suggests that the above difficulties in variance estimation and construction of confidence intervals arise from left-truncation.

Finally, we note that coverage probabilities are slightly worse when using the cross-validated tuning parameters. However, the difference is minimal, which provides reassurance about the reliability of the proposed data-driven choice of tuning values.

5. The statin study: Estimation of pregnancy outcome probabilities. The statin study aimed at estimating the risk of adverse pregnancy outcomes including spontaneous abortion associated with exposure to these drugs during pregnancy [Winterfeld et al. (2013)]. Statins are a class of drugs aiming at reducing cholesterol levels. Current guidelines advise to stop statin treatment during pregnancy, but since statins are widely prescribed and there is a trend for women to become pregnant later in life, an increasing number of women with childbearing potential

are likely to receive statin therapy, which might in turn increase the incidence of inadvertent fetal exposure. Details on the study can be found in Winterfeld et al. (2013).

Pregnancy outcomes of women, who had contacted a TIS on the use of statin during the first trimester of pregnancy were compared to a control group, which consisted of women seeking advice on drugs known to be non-teratogenic. Available for our data analysis were 235 women exposed to statin and 187 controls with complete information on study entry times and pregnancy outcomes.

The resulting CIFs for spontaneous and induced abortion and live birth of a first analysis using standard estimation techniques are displayed in Figure 5. Confidence intervals were computed using a Greenwood-type estimator as in (A.5) and a complementary log-minus-log transformation as in (4.1).

First of all, we notice that statin use during pregnancy increases the absolute risk of spontaneous abortion. Unexpectedly, however, standard estimation leads to a CIF of induced abortion of 0.37 at the plateau for the control group, which is also much higher than in the exposed group. Furthermore, this leads to an increased probability of live birth in the exposed group as compared to the controls, suggesting statin use to have a protective effect on both induced abortion and live birth.

As can be seen in the lower plot of Figure 5 this high estimand is due to the fact that an induced abortion happened in week 4 of the pregnancy when only 3 women were at risk in the control group, leading to a high variability and questionable estimates for the probability of induced abortion.

The practical implications of this were two-fold: From a statistical perspective, the high variability may possibly mask a difference between the curves because of loss of power. In our collaborative experience, however, discussion centered on the point estimate of induced abortion, which was considered to be implausible, and, as a consequence, the beneficial association (in terms of the point estimates) between use of statin and both induced abortion and live birth, which was medically unexpected.

A common solution of the problem is to estimate the CIFs conditional on being event-free up to week 4 of the pregnancy, that is, to estimate $P(T \leq t, X(T) = j | t > 4)$, $j = 1, 2, 3$. This was also the approach used by Winterfeld et al. (2013). However, this conditional estimation is ad hoc in the sense that the time point of 4 weeks is chosen based on the data at hand and not defined a priori. The advantage of the novel analyses below—using our modified Aalen–Johansen estimator—will be that it allows for estimating unconditional CIFs. This is relevant for our medical research question: Statin treatment is usually stopped during pregnancy, and the main aim of our analysis was to estimate the absolute outcome risks $P(X(T) = j)$, $j = 1, 2, 3$, for a statin user who becomes pregnant. Unlike the ad hoc solution, the new approach would also apply if further problematic time intervals had occurred after week 4, but still in the first trimester of pregnancy, in which our main interest lies.

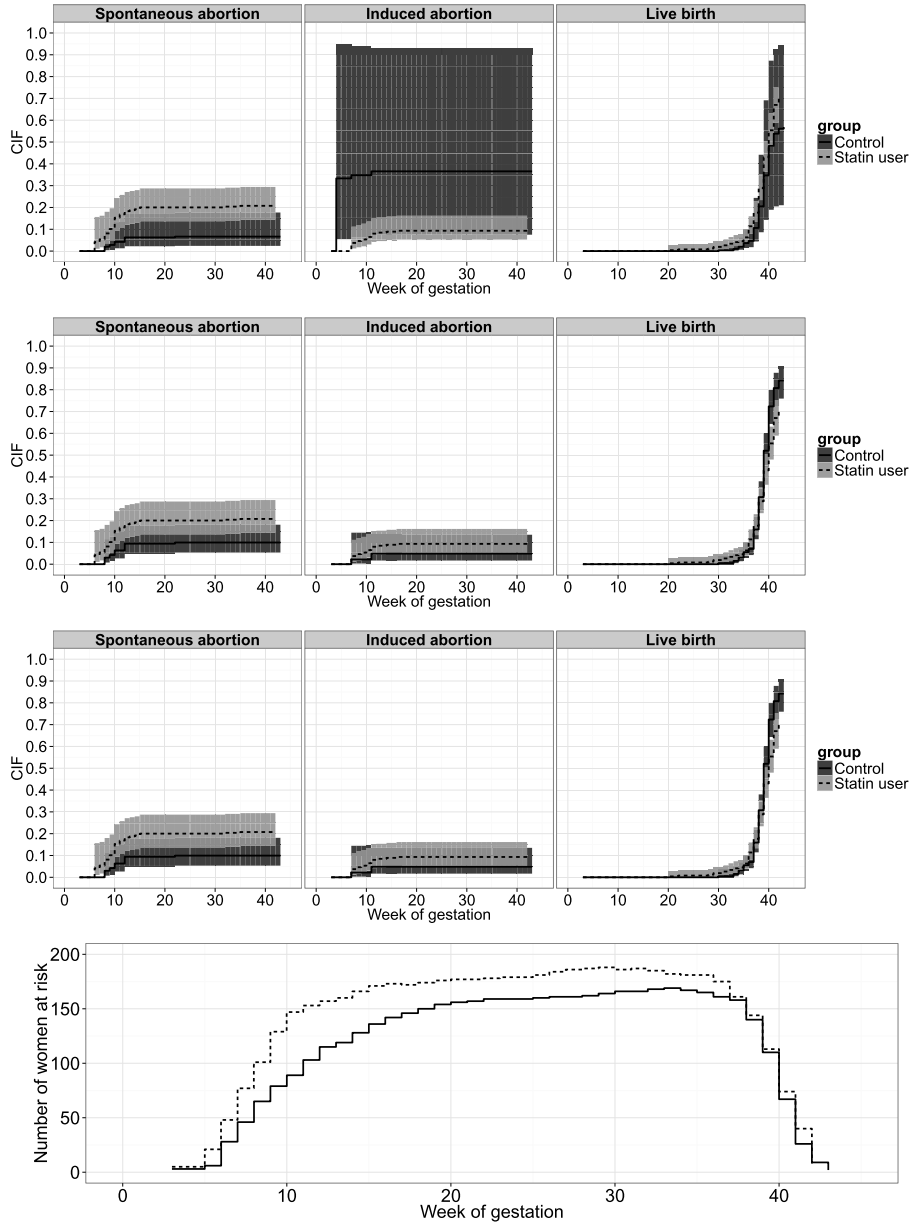

FIG. 5. *Results of the three analyses*: *The upper plot displays the CIFs for spontaneous abortion*, *induced abortion and live birth*, *estimated by the classical Aalen–Johansen estimator. The second plot shows the corresponding CIFs estimated by the modified Aalen–Johansen estimator using* $\lceil cn^{\gamma} \rceil = 4$. *The third plot shows the CIFs estimated by the modified Aalen–Johansen estimator using the results of the* 632 *bootstrap procedure*, *that is*, $\lceil cn_c^{\gamma} \rceil = 9$ *in the control group and* $\lceil cn_e^{\gamma} \rceil = 1$ *for the exposed. The lowest plot shows the number of women at risk at each time point in the control group and the exposed group*, *respectively.*

We start with a first reanalysis in Section 5.1, using $\gamma = 1/4$ and $c = 1$ as suggested by Lai and Ying (1991). Next, in Section 5.2, we cross-validate our choice of γ and c as in Section 4.3, again combined with a resampling approach in order to both train and apply the stabilized Aalen–Johansen estimator on the statin data. Finally, a simple two group comparison following the suggestion of Meister and Schaefer (2008) is in Section 5.3.

5.1. *A first reanalysis with* $\gamma = 1/4$ *and* $c = 1$. A first choice of $\gamma = 1/4$ and $c = 1$ implies a border of $\lceil cn^\gamma \rceil = 4$ for the exposed group ($n_e = 235$) as well as for the controls ($n_c = 187$). The resulting CIFs are displayed in Figure 5. Again, complementary log-minus-log transformed confidence intervals based on a Greenwood-type variance estimator are displayed.

Now, the CIF for induced abortion in the control group runs below the one for the exposed group and does not reach such an implausibly high plateau anymore. Moreover, the estimated probability for experiencing a live birth in the control group is now higher than for the statin users and confidence intervals are much smaller for both induced abortion and live birth.

These findings suggest that applying the modified Aalen–Johansen estimator is meaningful in this setting. However, we have seen in the simulation studies, that the choice of the tuning parameters is crucial for obtaining valid results. We would therefore like to determine optimal—data-driven—choices of c and γ for the analysis of this data.

5.2. *A cross-validated and resampled reanalysis*. In order to choose optimal parameters γ and c, we apply cross-validation combined with a resampling approach as in Section 4.3 in order to both train and apply the stabilized Aalen–Johansen estimator on the statin data. We applied the 632 bootstrap procedure described in Section 4.3 with $B = 10{,}000$ bootstrap runs to the pregnancy data, analyzing the two treatment groups separately.

We again considered values for the two parameters c and γ such that $\lceil cn^\gamma \rceil \in \{1, \ldots, 10\}$ for both the control ($n = n_c = 187$) and the exposed ($n = n_e = 235$) group. Note that the choices of the parameters for both groups coincide in most cases because sample sizes are comparable in the two groups.

Repeating the 632 bootstrap procedure for all choices of c and γ displayed in Table 3 and choosing the model with the smallest error $\widehat{\mathrm{Err}}_\omega$ resulted in $\lceil cn_c^\gamma \rceil = 9$ for the control group and $\lceil cn_e^\gamma \rceil = 1$ for the exposed. The results are displayed in Figure 5.

As we can see, there is no notable difference between the first reanalysis of the previous subsection and the present one. This is due to the fact that the risk set increases pretty fast after the first few weeks and minor changes in the bound cn^γ therefore do not alter the estimation much. However, the 632 bootstrap version is the preferred method since the choice of the tuning parameters is less arbitrary.

TABLE 3
Choices for c and γ such that $\lceil cn^{\gamma} \rceil \in \{1, \ldots, 10\}$ for $n_c = 187$ and $n_e = 235$, respectively

	Control $n = n_c = 187$		**Exposed $n = n_e = 235$**	
$\lceil cn^{\gamma} \rceil$	c	γ	c	γ
1	0.002	0.75	0.002	0.75
2	0.7	0.1	0.7	0.1
3	2	0.01	2	0.01
4	1	0.25	1	0.25
5	0.3	0.55	0.3	0.55
6	0.5	0.5	0.5	0.5
7	1.5	0.3	1.5	0.2
8	0.1	0.73	0.1	0.73
9	6	0.07	6	0.07
10	3.5	0.2	4	0.15

5.3. *Two group comparisons*. We have furthermore conducted a test in order to examine whether the probability of spontaneous or induced abortion at time $t = 40$ weeks differs between the two treatment groups.

Using the proposed estimator, the estimated cumulative incidence of spontaneous abortion by forty weeks was 0.21 (95% CI: [0.13, 0.28]) in the exposed group compared to 0.1 (95% CI: [0.04, 0.16]) in the control group. A two-sided comparison of these probabilities following Meister and Schaefer (2008) led to a p-value of 0.014 indicating a significant increase in the CIF of spontaneous abortion for the statin users. Using the classical Aalen–Johansen estimator led to the same results for the exposed group compared to 0.07 (95% CI: [0.00, 0.13]) in the control group and a p-value of 0.003.

For induced abortion, the proposed estimator led to an estimated cumulative incidence of 0.09 (95% CI: [0.04, 0.15]) in the exposed group compared to 0.05 (95% CI: [0.0, 0.1]) in the control group with a two-sided p-value of 0.114. Again, the classical Aalen–Johansen estimator led to the same results for the exposed group, but resulted in an estimate of 0.37 (95% CI: [−0.14, 0.87]) in the control group with a two-sided p-value of 0.852. While we cannot reject the null hypothesis for induced abortion for neither estimator, the p-value is much higher in the standard analysis, because the estimated probability in the control group is much higher with wider confidence intervals when using the classical Aalen–Johansen estimator. This again shows the high variability of the Aalen–Johansen estimator as a consequence of an event in an early small risk set.

6. Discussion. We have developed a stabilized Aalen–Johansen estimator for the cumulative event probability of a competing risk that discards contributions

from risk sets that are too small to produce reliable estimates. The motivation was a study on the use of statin during pregnancy. The standard Aalen–Johansen estimator does account for delayed study entries of pregnant women, but produced a medically unexpected finding because of an early induced abortion in an overly small risk set. This has been remedied by the new approach. We first discuss aspects which are more closely linked to the data application and consider more general aspects afterward.

To begin, we have argued that, in terms of the point estimates, the beneficial association between use of statin and both induced abortion and live birth was medically unexpected or even implausible. However, several potential confounders that could not be controlled for in the analysis may have influenced the induced abortion of pregnancy rates in the statin-exposed and the control group. Factors that might have increased the risk in the statin group include concerns of pregnant women or their physicians regarding underlying disease as well as inadvertent exposure in the beginning of pregnancy and their consequences on pregnancy outcome. Since women are currently advised to discontinue statin treatment when planning a pregnancy, there also might have been a higher rate of unplanned pregnancies with poorer acceptance in the statin group. In contrast, socio-economic status might have differed between groups with affluent women potentially taking more statins while also terminating pregnancies less often.

Next, we have argued that the conditional analysis of Winterfeld et al. (2013) was meaningful, but ad hoc, because the time point of conditioning had not been specified in advance. We have also argued that the original research question would be addressed best by an unconditional analysis. Two remarks are in place: First, there will be research questions where conditional analyses may be most adequate. In the current context, one may envisage interest in preterm birth conditional on survival of the first trimester of pregnancy. Such research might also be more interested in the actual timing of events than our present analysis was. Second, an analysis based on the stabilized Aalen–Johansen estimator will be ad hoc, too, if the tuning parameters for formalizing an overly small risk set are chosen *ad libitum*.

We have used cross-validation for selection of the tuning parameters, combined with a 632 bootstrap procedure in order to both train and apply our procedure on one data set. In the process of cross-validation, one has to make a decision as to which measure of (prediction) error one has to optimize. We have suggested to use an estimated Brier Score for left-truncated pregnancy outcomes. This involves some characteristics for the data problem at hand: Major interest lies in the eventual outcome type, and there is less interest in its timing. In the cross-validation, we have used updated prediction given time of study entry. Because once under study, pregnancy outcomes are observed, there was no need for Inverse Probability of Censoring Weighting techniques, but the expected Brier Score is w.r.t. a probability measure given study entry.

We have made an independent left-truncation assumption as in Andersen et al. (1993) that assumes that the compensator of the observable counting process coincides with the compensator in the absence of left-truncation save for the modified at-risk status. A concern is that this may be violated for induced abortion outcomes. As stated in the Introduction, one aim of TIS counseling is to reduce the rate of induced abortions based on irrational overestimation of drug risks. On the other hand, Allignol, Schumacher and Beyersmann (2010) found in a larger study on the effect of coumarin anticoagulants during pregnancy that study entry may have an effect on the hazard of induced abortion. The issue is not at all clear: Testing independent left-truncation has mainly been investigated for all-cause survival outcomes, for example, Tsai (1990). Competing risks complicate the situation. It is possible that study entry has an effect on the cause-specific hazards for events 2 and 3, but not on their sum, which would be the competing hazard for event 1. The issue merits further research, possibly using a synthesis model as in Beyersmann and Schumacher (2008). It will also be worthwhile to study in more detail modeling of dependent left-truncation as in Mackenzie (2012), but in the presence of competing risks. In either case, delayed study entry must be accounted for and, therefore, the present approach improves on the commonly used multinomial estimates.

As stated in Section 4.2, the new approach avoids implausibly large point estimates. As a consequence, it tends to display a larger absolute mean bias, but median biases are comparable between the modified and the standard Aalen–Johansen estimator. Another consequence is a smaller variation resulting in smaller widths of the confidence intervals. Because the coverage probabilities are comparable, we prefer the new procedure. However, the coverage probabilities for both procedures leave something to be desired and merit future research, possibly using resampling procedures.

A major extension along the lines of the present work will be to study the Aalen–Johansen estimator of the transition matrix of a time-inhomogeneous Markov process with finite state space. Unlike Lai and Ying, we have chosen to first work with the multivariate Nelson–Aalen estimator and to subsequently translate results for probability estimation using product integration. Therefore, the generalization should be technically straightforward, but we expect it to be also relevant for such multistate models even in the absence of left-truncation: For instance, in an illness-death model without recovery, initial state 0, intermediate illness state 1 and absorbing death state 2, there will be *internal* left-truncation due to $0 \to 1$ transitions. Examples with unstable estimation can be constructed along the lines of Section 4, even if there are no delayed study entries and all individuals are prospectively followed-up starting in state 0 at time 0.

Three final technical remarks are in place: First, we have used martingale arguments throughout, using Andersen et al. (1993) as a template. This has substantially simplified our proofs compared to Lai and Ying (1991). However, two technical ideas of Lai and Ying that are not used in the theory of Andersen et al. were

important in our developments. First, we viewed the actual sample under study of size n as being generated by a larger sample $m(n)$. This mirrors the common simulation approach of simulating $m(n)$ individuals, but only treating n, $n < m(n)$, individuals as being observed; see Section 4. The nice technical consequence is that one does not have to work with conditional probability measures given study entry. (An alternative simulation approach would be to simulate study entries and to subsequently simulate event times.) Second, we did not work with the usual filtration where the past encodes all data observed by the researcher so far, but we followed the approach of Lai and Ying and enlarged the usual filtration with additional knowledge of the study entry times and under study statuses of all $m(n)$ individuals. Technically, this ensured predictability of the bound cn^γ, and hence, enabled use of martingale methods. Interpretationally, this is potentially awkward because knowledge of *future* study entry is unknown in practice. However, for most applications this can be viewed as a mere technical device that ensures predictability, but does not compromise practical inference. The assumption is that the momentary intensity of an individual under study does not depend on future study entries of other individuals. However, counter examples may be constructed (see Remark 7.1 in the supplement [Friedrich et al. (2017)]) and, therefore, our assumptions are more restrictive than those of Andersen et al. (1993).

APPENDIX: TECHNICAL DETAILS AND PROOFS

This Appendix contains the proofs of Theorem 3.2 and Theorem 3.3 as well as a section on variance estimation.

PROOF OF THEOREM 3.2. Consider the modified estimator:

$$\hat{A}_{0j}(t) = \int_0^t \frac{\mathbb{1}(Y(s) \geq cn^\gamma)}{Y(s)}\, dN_{0j}(s)$$

and let

$$A^*_{0j}(t) = \int_0^t \mathbb{1}\big(Y(s) \geq cn^\gamma\big)\alpha_{0j}(s)\, ds. \tag{A.1}$$

Then

$$H(s) = \frac{\mathbb{1}(Y(s) \geq cn^\gamma)}{Y(s)} \tag{A.2}$$

is predictable w.r.t. $\mathcal{F}(s)$, since $Y(s)$ and n are known at $s-$. Introduce the mean-zero martingale:

$$Z_{0j}(t) = \int_0^t H(s)\, dM_{0j}(s) = \hat{A}_{0j}(t) - A^*_{0j}(t), \qquad j = 1, 2 \tag{A.3}$$

with predictable covariation process given by

$$\begin{aligned}\langle Z_{0j}, Z_{0l}\rangle(t) &= \left\langle \int_0^t H(s)\,dM_{0j}(s), \int_0^t H(s)\,dM_{0l}(s)\right\rangle \\ &= \int_0^t H^2(s)\,d\langle M_{0j}, M_{0l}\rangle(s) \\ &= \delta_{jl}\int_0^t H(s)\alpha_{0j}(s)\,ds.\end{aligned}$$

Here, δ_{jl} denotes the Kronecker delta, that is, $\delta_{jl}=1$ for $j=l$ and 0, otherwise.

The martingales M_{01} and M_{02} are orthogonal, since the new filtration—as explained above—leaves the intensity processes of the martingales unchanged.

With Lenglart's inequality [Andersen et al. (1993), Section II.5.2.1], we get for any $\delta, \eta > 0$:

$$P\Big(\sup_{s\in[0,t]}\big|\hat{A}_{0j}(s) - A^*_{0j}(s)\big| > \eta\Big) \le \frac{\delta}{\eta^2} + P\bigg(\int_0^t \frac{\mathbb{1}(Y(s)\ge cn^\gamma)}{Y(s)}\alpha_{0j}(s)\,ds > \delta\bigg).$$

By (3.4), it follows that

$$\sup_{s\in[0,t]}\big|\hat{A}_{0j}(s) - A^*_{0j}(s)\big| \xrightarrow{P} 0, \qquad n\to\infty.$$

Furthermore,

$$\big|A^*_{0j}(s) - A_{0j}(s)\big| = \int_0^s \big(1-\mathbb{1}\big(Y(u)\ge cn^\gamma\big)\big)\alpha_{0j}(u)\,du \xrightarrow{P} 0$$

by (3.5) and consistency of the modified estimator follows.

The final conclusion follows directly by applying the max-supremum norm. □

PROOF OF THEOREM 3.3. Recall from (A.3)

$$Z_{0j}(t) = \int_0^t H(s)\,dM_{0j}(s) = \hat{A}_{0j}(t) - A^*_{0j}(t), \qquad j=1,2,$$

that is, Rebolledo's martingale central limit theorem [Andersen et al. (1993), Theorem II.5.1] applies. By Conditions (1) and (2), it follows immediately that

$$\sqrt{n}\big(\hat{\mathbf{A}} - \mathbf{A}^*\big) \xrightarrow{\mathcal{D}} \mathbf{W} \qquad \text{as } n\to\infty.$$

Furthermore, Condition (3) assures

$$\sup_{s\in[0,t]}\sqrt{n}\big|A^*_{0j}(s) - A_{0j}(s)\big| \xrightarrow{P} 0$$

as $n\to\infty$ and the conclusion follows. □

A.1. Variance estimation. Lai and Ying (1991) did not provide variance estimators. We derive such estimators using martingale theory.

A.1.1. *Nelson–Aalen estimator.* The aim is to estimate the variance $\sigma_j^2(t)$. Therefore, consider the optional variation process of $\hat{A}_{0j} - A_{0j}^*$ for $j = 1, 2$:

$$\begin{aligned}\hat{\sigma}_j^2(t) &= [\hat{A}_{0j} - A_{0j}^*](t) = [Z_{0j}](t) \\ &= \int_0^t H^2(s)\, d[M_{0j}](s) = \int_0^t \frac{\mathbb{1}(Y(s) \geq cn^\gamma)}{Y^2(s)}\, dN_{0j}(s) \\ &= \sum_{s \leq t} \frac{\Delta N_{0j}(s)}{Y^2(s)} \mathbb{1}\big(Y(s) \geq cn^\gamma\big).\end{aligned} \tag{A.4}$$

Rebolledo's martingale central limit theorem now ensures that the covariance $\sigma_j^2(t)$ of the Gaussian martingale $\mathbf{W}$ from Theorem 3.3 may be consistently estimated by $n \cdot \hat{\sigma}_j^2(t)$.

A.1.2. *Aalen–Johansen estimator.* Using the same arguments as in Andersen et al. (1993), Chapter IV.4, an estimator for the covariance of the modified Aalen–Johansen estimator is given by

$$\widehat{\mathrm{Cov}}\big(\hat{\mathbf{P}}(s,t)\big) = \int_s^t \hat{\mathbf{P}}(u,t)^\top \otimes \hat{\mathbf{P}}(s,u)\, d\big[\hat{\mathbf{A}} - \mathbf{A}^*\big](u) \hat{\mathbf{P}}(u,t) \otimes \hat{\mathbf{P}}(s,u)^\top,$$

where $\hat{\mathbf{P}}(s,t)$ denotes the modified Aalen–Johansen estimator, $[\hat{\mathbf{A}} - \mathbf{A}^*](t)$ is given by (A.4) and $\otimes$ is the Kronecker product of matrices. For competing risks, the variance of the estimated cumulative incidence function may be estimated by a Greenwood-type estimator:

$$\begin{aligned}\widehat{\mathrm{var}}\big(\hat{P}_{0j}(0,t)\big) &= \sum_{s \leq t} \frac{(\hat{P}_{0j}(0,t) - \hat{P}_{0j}(0,s))^2}{Y(s) - \Delta N_{0\cdot}(s)} \mathbb{1}\big(Y(s) \geq cn^\gamma\big) \Delta \hat{A}_{0\cdot}(s) \\ &\quad + \frac{\hat{S}(s-)^2}{Y(s)^3} \cdot \mathbb{1}\big(Y(s) \geq cn^\gamma\big) \cdot \Big[Y(s) - \Delta N_{0j}(s) \\ &\quad - 2\big(Y(s) - \Delta N_{0\cdot}(s)\big) \cdot \frac{\hat{P}_{0j}(0,t) - \hat{P}_{0j}(0,s)}{\hat{S}(s)} \Big] \Delta N_{0j}(s),\end{aligned} \tag{A.5}$$

analogous to Allignol, Schumacher and Beyesmann (2010), equation (6), but using the modified estimators of Section 3. This is also how the variance estimators in Sections 4 and 5 have been implemented.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric estimation of pregnancy outcome probabilities" (DOI: 10.1214/17-AOAS1020SUPP; .pdf). We discuss the classical survival case and a modified Kaplan–Meier estimator and provide additional simulation results.

REFERENCES

AALEN, O. O. and JOHANSEN, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Stat.* **5** 141–150. MR0509450

ALLIGNOL, A., SCHUMACHER, M. and BEYERSMANN, J. (2010). A note on variance estimation of the Aalen-Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biom. J.* **52** 126–137. MR2756598

ANDERSEN, P. K. and KEIDING, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Stat. Med.* **31** 1074–1088. MR2925679

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

ANDERSEN, A.-M. N., ANDERSEN, P. K., OLSEN, J., GRØNBÆK, M. and STRANDBERG-LARSEN, K. (2012). Moderate alcohol intake during pregnancy and risk of fetal death. *Int. J. Epidemiol.* **41** 405–413.

BEYERSMANN, J., ALLIGNOL, A. and SCHUMACHER, M. (2012). *Competing Risks and Multistate Models with R*. Springer, New York. MR3025354

BEYERSMANN, J. and SCHUMACHER, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics* **9** 765–776.

BEYERSMANN, J., LATOUCHE, A., BUCHHOLZ, A. and SCHUMACHER, M. (2009). Simulating competing risks data in survival analysis. *Stat. Med.* **28** 956–971. MR2518359

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331. MR0711106

EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560. MR1467848

FRIEDRICH, S., BEYERSMANN, J., WINTERFELD, U., SCHUMACHER, M. and ALLIGNOL, A. (2017). Supplement to "Nonparametric estimation of pregnancy outcome probabilities." DOI:10.1214/17-AOAS1020SUPP.

GERDS, T. A. and SCHUMACHER, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48** 1029–1040. MR2312613

GERDS, T. A. and SCHUMACHER, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63** 1283–1287, 1316. MR2414608

GOOLEY, T. A., LEISENRING, W., CROWLEY, J., STORER, B. E. et al. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Stat. Med.* **18** 695–706.

GRZESKOWIAK, L. E., GILBERT, A. L. and MORRISON, J. L. (2012). Exposed or not exposed? Exploring exposure classification in studies using administrative data to investigate outcomes following medication use during pregnancy. *Eur. J. Clin. Pharmacol.* **68** 459–67.

HANCOCK, R. L., KOREN, G., EINARSON, A. and UNGAR, W. J. (2007). The effectiveness of teratology information services (TIS). *Reprod. Toxicol. (Elmsford N.Y.)* **23** 125–132.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

HELD, L. and SABANÉS BOVÉ, D. (2014). *Applied Statistical Inference*. Springer, Heidelberg. MR3155114

KEIDING, N. and GILL, R. D. (1990). Random truncation models and Markov processes. *Ann. Statist.* **18** 582–602. MR1056328

KLEIN, J. P. (1991). Small sample moments of some estimators of the variance of the Kaplan–Meier and Nelson–Aalen estimators. *Scand. J. Stat.* **18** 333–340. MR1157787

LAI, T. L. and YING, Z. (1991). Estimating a distribution function with truncated and censored data. *Ann. Statist.* **19** 417–442. MR1091860

LUPATTELLI, A., SPIGSET, O., TWIGG, M. J. et al. (2014). Medication use in pregnancy: A cross-sectional, multinational web-based study. *BMJ Open* **4**.

MACKENZIE, T. (2012). Survival curve estimation with dependent left truncated data using Cox's model. *Int. J. Biostat.* **8** Art. 29. MR2997679

MEISTER, R. and SCHAEFER, C. (2008). Statistical methods for estimating the probability of spontaneous abortion in observational studies—analyzing pregnancies exposed to coumarin derivatives. *Reprod. Toxicol.* (*Elmsford N.Y.*) **26** 31–35.

SCHOOP, R., BEYERSMANN, J., SCHUMACHER, M. and BINDER, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biom. J.* **53** 88–112. MR2767380

SLAMA, R., BALLESTER, F., CASAS, M., CORDIER, S., EGGESBØ, M., INIGUEZ, C., NIEUWENHUIJSEN, M., PHILIPPAT, C., REY, S., VANDENTORREN, S. et al. (2014). Epidemiologic tools to study the influence of environmental factors on fecundity and pregnancy-related outcomes. *Epidemiol. Rev.* **36** 148–164.

TSAI, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77** 169–177. MR1049418

VAN HOUWELINGEN, H. C. and PUTTER, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. *Monographs on Statistics and Applied Probability* **123**. CRC Press, Boca Raton, FL. MR3058205

WILLAND, I. (2011). *Statistisches Jahrbuch*. Statistisches Bundesamt, Wiesbaden.

WILLAND, I. (2014). *Statistisches Jahrbuch*. Statistisches Bundesamt, Wiesbaden.

WINTERFELD, U., ALLIGNOL, A., PANCHAUD, A., ROTHUIZEN, L. E., MERLOB, P., CUPPERS-MAARSCHALKERWEERD, B., VIAL, T., STEPHENS, S., CLEMENTI, M., SANTIS, M. et al. (2013). Pregnancy outcome following maternal exposure to statins: A multicentre prospective study. *BJOG*: *An International Journal of Obstetrics and Gynaecology* **120** 463–471.

S. FRIEDRICH
J. BEYERSMANN
A. ALLIGNOL
INSTITUTE OF STATISTICS
ULM UNIVERSITY
ULM
GERMANY
E-MAIL: sarah.friedrich@uni-ulm.de
jan.beyersmann@uni-ulm.de
arthur.allignol@uni-ulm.de

U. WINTERFELD
STIS AND DIVISION OF CLINICAL PHARMACOLOGY AND TOXICOLOGY
UNIVERSITY HOSPITAL LAUSANNE
LAUSANNE
SWITZERLAND
E-MAIL: ursula.winterfeld@chuv.ch

M. SCHUMACHER
INSTITUTE OF MEDICAL BIOMETRY AND MEDICAL INFORMATICS
UNIVERSITY MEDICAL CENTER FREIBURG
FREIBURG
GERMANY
E-MAIL: ms@imbi.uni-freiburg.de