STATISTICA, anno LXXX, n. 4, 2020

VARIANCE INFLATION DUE TO CENSORING IN SURVIVAL PROBABILITY ESTIMATES

Szilárd Nemes [1]
BioPharma Early Biometrics and Statistical Innovation, Data Science & AI, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

Andreas Gustavsson
BioPharma Early Biometrics and Statistical Innovation, Data Science & AI, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

Ziad Taib
Atlasbiostat, Gothenburg, Sweden

1. Introduction

Since the seminal work by Kaplan and Meier (1958), the product limit estimator (or Kaplan-Meier curve) had become the main (almost single) way of summarizing survival data. The Kaplan Meier curve is a non-parametric maximum likelihood estimator of survival probability with efficacy comparable to the more sophisticated parametric models (Meier *et al.*, 2004). The main attractiveness of the Kaplan-Meier curve is the simple but elegant way of handling incomplete data, a common feature of survival analysis. This partial loss of information due to censoring ought to inflate the variance of the estimated probabilities. Functionals of the Kaplan-Meier curve, namely the restricted mean survival time (Akritas, 2000; Stute, 1995a,b, 2003) or life expectancy (Yang, 1977) received a fair share of interest, regarding the issue of the variance under censoring. As far as we know Cantor (2001) was the first to study the issue of variance inflation of Kaplan-Meier curves under censoring. Prior to that, Brooks (1982) considered the information loss for exponential survival times due to censoring while Zheng and Gastwirth (2001) derived the Fisher information under censoring for parameter estimates. The relative inattention towards this issue is likely to be due to the existence of standard routines for statistical inference that can be applied to survival curves, some of which predates the Kaplan-Meier curve, such as the Greenwood variance estimator (Greenwood, 1926) or Irwin's accrual estimates (Irwin, 1949).

[1] Corresponding Author. E-mail: szilard.nemes@astrazeneca.com

In this paper we derive a simple estimator for the variance inflation due to incomplete observations. This estimator allows the estimation of the expected variance under different assumptions for survival and censoring times. The properties of the proposed estimator are illustrated using simulation studies and its applicability is highlighted based on published data.

2. NOTATION

We assume that survival times to an event of interest $X_1,\ldots,X_n$ to be independently and identically distributed according to the distribution function $F(t)$ and survival function $S(t)=1-F(t)=\exp\left\{-\int_0^t \lambda(u)\,du\right\}$, where $\lambda(t)$ is the hazard function for the event. However, we do not have information for all subjects due to loss to follow-up or insufficient follow-up time. The times to censoring are denoted by $C_1,\ldots C_n$, and assumed to be independently and identically distributed according to the distribution function $C(t)$ and survival function $G(t)=\exp\left\{-\int_0^t \gamma(u)\,du\right\}$ where $\gamma(t)$ is the hazard function for censoring. Thus the actual observed time for subject j is $T_j=\min\left(X_j,\,C_j\right)$. Furthermore, we assume independence between failure and censoring time.

Additionally, we define $\delta_j=I\left\{X_j\le C_j\right\}$ as an event indicator, taking values 1 if an event is recorded prior to censoring, otherwise 0. We define $N(t)$ to be the counting process that counts the number of individuals experiencing the event in the interval $[0,\,t]$. For a right censored subject $N_j(t)=I\left(T_j\le t,\delta_j=1\right)$. The aggregated process is given by $N(t)=\sum_{j=1}^n N_j(t)=\sum_{t_j\le t}\delta_j$. $N_j(t)$ is a right-continuous piecewise constant function with jumps of size 1. The jumps $dN(t)=N(t+dt)^-N(t)^-$ are changes in the process over a short interval $[t\;+dt)$. Similarly we define the number at risk at time t as $Y(t)=\sum_{j=1}^n I\left(T_j\ge t\right)$. $Y(t)$ is a decreasing step function with steps of size 1 at each event or censoring time.

3. VARIANCE INFLATION OF THE KAPLAN-MEIER ESTIMATE

When survival times are censored the distribution function for the survival time F can be estimated by the Kaplan-Meier estimator $\hat{F}$, commonly expressed as the survival function

$$\hat{S}(t)=\prod_{T_i\le t}\left[1-\frac{dN(T_i)}{Y(T_i)}\right], \tag{1}$$

where T_i is the ordered sequence of event times. In case of no censoring, the variance of the survival function should equal the variance of a binomially distributed variable, i.e. $\hat{S}(t)\left(1-\hat{S}(t)\right)/n$ (Meier, 1975; Aalen *et al.*, 2008).

Several expressions for the variance of $\hat{S}(t)$ are available when censoring is present. Among these, the Greenwood estimator is well known and most used. According to the Greenwood estimator the standard deviation is given by $\sqrt{\widehat{\text{var}}(S(t))}$, where

$$\widehat{\text{var}}(S(t)) = \hat{S}^2(t) \sum_{T_i \le t} \frac{dN(T_i)}{Y(T_i)[Y(T_i) - dN(T_i)]}. \tag{2}$$

If the exact survival time is unknown for some subjects, i.e. censoring is present then $\widehat{\text{var}}(S(t)) > \hat{S}(t)\left(1 - \hat{S}(t)\right) n^{-1}$.

We let the time dependent **variance inflation** factor $\varphi(t)$ quantify how much the variability of a survival probability estimate increases due to censoring.

Assuming a multiplicative effect of censoring on the estimated variance we postulate that

$$\text{var}(S(t)) = \varphi(t) S(t)(1 - S(t)) n^{-1}. \tag{3}$$

Derivation of $\varphi(t)$ proceeds with division of the observed follow-up times T_i into m equally sized intervals $\Delta y = y_i - y_{i-1} = m^{-1} t$ on $[0,\ t]$ as $\{0 < y_1 < \ldots < y_m = t\}$. The number of events recorded in each time interval depends on the numbers at risk, the hazard and the length of the interval, $dN(y_i) = \lambda(y_i) Y(y_i) \Delta y$. Additionally, we can express the number of patients at risk as the expectation of a binomial distribution with parameters, n the sample size and the probability of being at risk given by the product of probabilities of not experiencing the event and the probability of not being censored. Thus, $Y(y_i) = n S(y_i) G(y_i)$ (Andersen *et al.*, 2012).

We can rewrite

$$\sum_{y_i \le t} \frac{dN(y_i)}{Y(y_i)[Y(y_i) - dN(y_i)]} = \sum_{y_i \le t} \frac{\lambda(y_i) Y(y_i) \Delta y}{Y(y_i)[Y(y_i) - \lambda(y_i) Y(y_i) \Delta y]}.$$

This can be simplified to

$$\sum_{y_i \le t} \frac{\lambda(y_i) \Delta y}{[Y(y_i) - \lambda(y_i) Y(y_i) \Delta y]}.$$

If we increase the number of intervals $m \to \infty$ then $\Delta y \to 0$ and

$$\lim_{\Delta y \to 0} \sum_{y_i \le t} \frac{\lambda(y_i) \Delta y}{[Y(y_i) - \lambda(y_i) Y(y_i) \Delta y]} = n^{-1} \int_0^t \frac{\lambda(u)}{1 - F(u) G(u)} du. \tag{4}$$

We can see that due to Eq. (3)

$$\varphi(t)\{1 - S(t)\} = S(t) \int_0^t \frac{\lambda(u)}{S(u) G(u)} du$$

and thereby

$$\varphi(t) = \frac{S(t)}{1-S(t)} \int_0^t \frac{\lambda(u)}{S(u)G(u)} du. \tag{5}$$

REMARK 1. *If there is no censorship then* $\varphi(t) = 1 \; \forall t > 0$.

PROOF. If there is no censorship the Greenwood estimator is simplifies to a binomial variance thus $\mathrm{var}(S(t)) = S(t)(1-S(t))n^{-1}$. This equality holds if and only if $\varphi(t) = 1$. □

REMARK 2. *If there is censorship (i.e.* $C(t) = P(C < t) > 0$*), then* $\varphi(t) > 1$.

PROOF. $S(t) \in (0,1)$ and $G(t) \in (0,1)$ thus $G(t)S(t) < S(t) \Rightarrow \lambda(t)(S(t)G(t))^{-1} > \lambda(t)S(t)^{-1}$ and $\int_0^t \frac{\lambda(u)}{(S(u)G(u))} du > \int_0^t \frac{\lambda(u)}{(S(u))} du \Rightarrow \phi(t) > 1$. □

REMARK 3. *Median survival time is often of special interest in research planing/analysis of data, at* $S(t) = 0.5$ *the equation simplifies to*

$$\varphi(t) = 2 \int_0^{t_{0.5}} \lambda(u)G(u)^{-1} du. \tag{6}$$

REMARK 4. *Let* $G(t)$ *be the assumed survival function for the censoring and* $G^*(t)$ *is the true survival function for the censoring. Then the difference between the projected* ($\varphi(t)$) *and true* ($\varphi^*(t)$) *variance inflation is given by*

$$\Delta_{\varphi(t),\varphi^*(t)} = \frac{S(t)}{1-S(t)} \int_0^t \frac{\lambda(u)}{S(u)} \left[\frac{1}{G(u)} - \frac{1}{G^*(u)} \right] du.$$

From Remark 4 it follows that

1. if $G(t) \equiv G^*(t)$ then $\Delta_{\varphi(t),\varphi^*(t)} = 0$;
2. if $G(t) < G^*(t)$ then $\Delta_{\varphi(t),\varphi^*(t)} > 0$;
3. if $G(t) > G^*(t)$ then $\Delta_{\varphi(t),\varphi^*(t)} < 0$.

In practical applications it may be convenient to provide the deviation from the expectation on a percentage scale (i.e. percent bias) (Burton *et al.*, 2006) which can be assessed numerically by $\Delta_{\varphi(t),\varphi^*(t)}/\varphi^*(t) \times 100$.

4. SIMULATION STUDIES

4.1. Exponential distribution

The distribution for the survival time and censoring time can take various forms. For most distributions closed form solutions for $\varphi(t)$ are difficult or impossible to obtain. If we assume that both X and C are exponentially distributed with survival functions $S(t)$ and $G(t)$ and hazard for event λ and hazard for censoring, γ then the variance inflation due censoring can be expressed as

$$\varphi(t) = \frac{f(t)}{F(t)} \int_0^t \frac{1}{S(u)G(u)} du = \frac{\lambda}{\lambda+\gamma} \frac{\left(e^{t(\lambda+\gamma)}-1\right)}{\left(e^{t\lambda}-1\right)}. \tag{7}$$

This estimator has some interesting intrinsic properties. The first component $\lambda(\lambda+\gamma)^{-1}$ is an estimate for $P(X<C) = \int_0^t C(u) f(u)\, du$ (Nadarajah, 2003) and the asymptotic relative efficiency (ARE) of the estimate of the rate parameter under censoring vs non-censoring. The second component is the ratio of odds recording and event within time t when censoring is present, $\left(e^{t(\lambda+\gamma)}-1\right)$ and the odds recording and event within time t when no censoring is present $\left(e^{t\lambda}-1\right)$.

Figure 1 summarizes the results of the simulation study. Assuming exponential distribution for both the event of interest ($\lambda = 1/365$) and censoring ($\gamma = 1/365$) we simulated 1000 observations. A survival curve was fitted to the data and its variability was estimated with the Greenwood formula.

If there would be no censoring the variance of the survival function would be $e^{-t\lambda}(1-e^{-t\lambda})n^{-1}$. Under censoring the expected variance is given by $\varphi(t)e^{-t\lambda}(1-e^{-t\lambda})n^{-1}$. Bias was defined as the projected standard error minus the standard error returned by the Greenwood formula. We iterated this procedure 10^4 times. If there is no censoring the standard error of a survival curve reaches its maximum at the median survival time ($S(t) = 0.5$) thereafter decreases again. Given the assumed hazard ($\lambda = 1/365$) and exponential distribution this corresponds to 252 days. As shown on Figures 1a and 1b, the estimated SE of the survival curves does not decrease after 252 days but monotonically increases with time, in concordance with the increase in variance inflation due to censoring. The projected SE ($\varphi(t)e^{-t\lambda}(1-e^{-t\lambda})n^{-1}$) coincides well with the SE estimated by the Greenwood estimator (Figure 1c). The empirical bias is negative and it increases with time (Figure 1d). However this bias is negligible compared to the estimated SE. The bias is roughly 700 times smaller than the estimated SE (range: 300-1800).

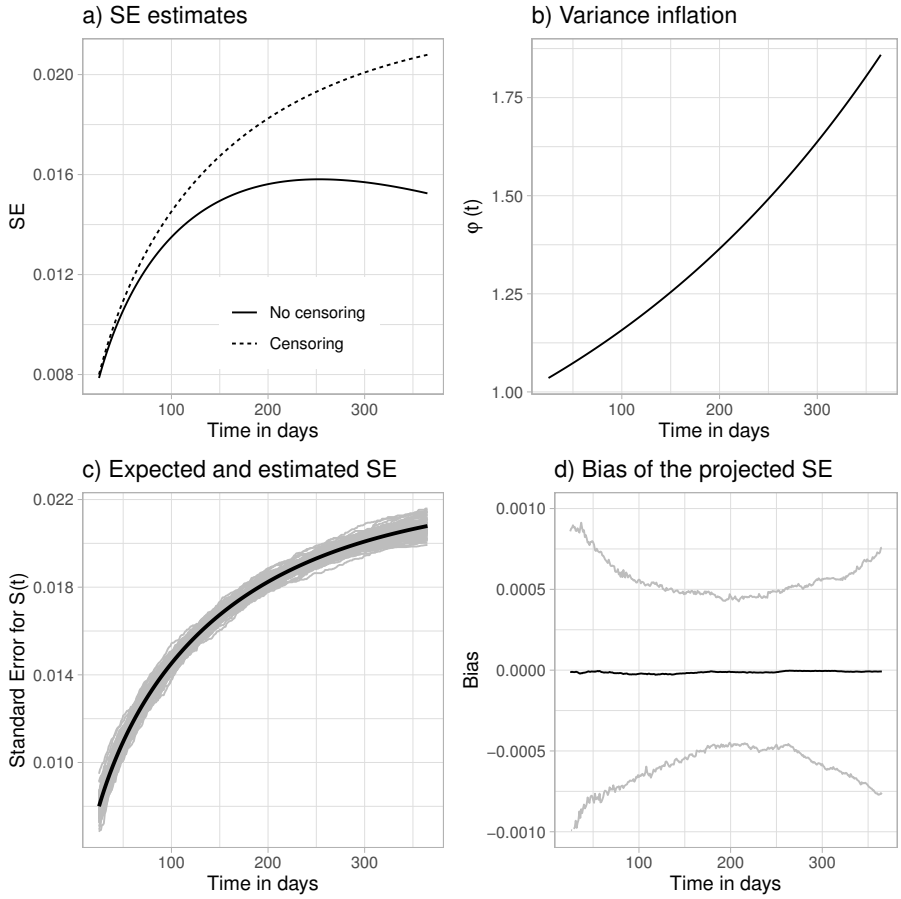

Figure 1 – Results of a simulation study assuming exponential survival and censoring times depicting (a) the expected SE for the survival curves with and without censoring, (b) the time dependent variance inflation due to censoring, (c) the expected (thick black) and estimated (thin grey lines) SE of the survival curve and (d) the bias of the projected SE with 95 % Monte Carlo confidence intervals.

4.2. Accelerated failures

In many practical applications failure rates are not constant over time but increasing (or decreasing) with time. The Weibull and log-logistic distributions are commonly used for modelling accelerated failure models.

The Weibull distribution can be described by two parameters, shape ($k > 0$) and scale ($\theta > 0$), with survival and hazard function given by

$$S(t) = \exp\left\{-\left(\frac{t}{\theta}\right)^k\right\} \quad \text{and} \quad \lambda(t) = \frac{k}{\theta}\left(\frac{t}{\theta}\right)^{k-1}.$$

If the scale parameter $\theta > 1$ the failure rate is increasing with time, a natural ageing process. Here we assumed the survival time follows the Weibull distribution with scale of 3 and shape of 365. We assumed memoryless/exponential censoring ($\gamma = 1/100$). For Weibull survival times and exponential censoring, the variance inflation factor is given by

$$\varphi(t) = \frac{e^{-(t/\theta)^k}}{1 - e^{-(t/\theta)^k}} \frac{k}{\theta} \int_0^t \frac{(u\theta^{-1})^{k-1}}{e^{-(u/\theta)^k} e^{-u\gamma}}\, du.$$

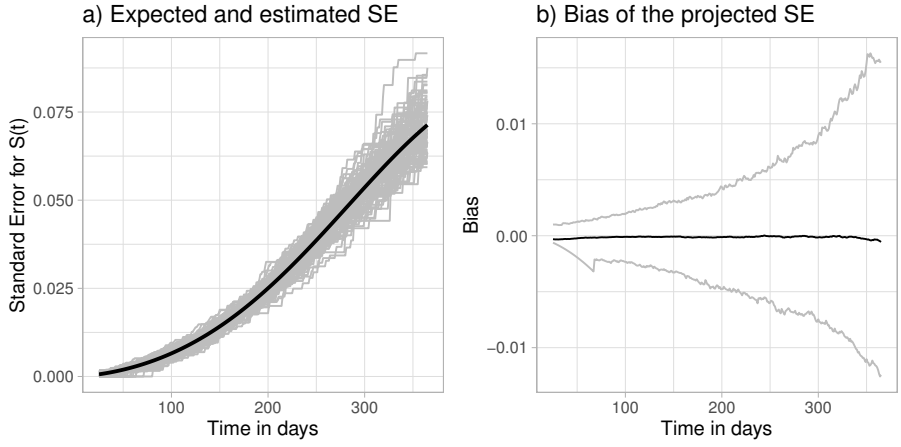

Figure 2 – Results of a simulation study assuming Weibull survival and exponential censoring times depicting (a) the expected (thick black) and estimated (thin grey lines) SE of the survival curve and (b) the bias of the projected SE with 95 % Monte Carlo confidence intervals.

Figure 2 summarizes the results of 10^4 simulations with the above described parameters. At the very beginning of the follow up period we observed a negative bias, however the bias was small compared to the variance.

The log-logistic is another distribution commonly used in accelerated failures models, described by the scale ($\alpha > 0$) and shape ($\beta > 0$) parameters with survival and hazard function given by

$$S(t) = \frac{1}{1+(t/\alpha)^{\beta}} \quad \text{and} \quad \lambda(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1+(t/\alpha)^{\beta}}.$$

Unlike the Weibull distribution, the hazard of the log-logistic distribution can take non-monotonic forms. The form of the hazard is governed by the shape parameter. We assumed that $\beta = 1.5$ resulting in a quadratic like hazard that increases up to 200 days and thereafter decreases. The scale parameter was set to 500 giving a mean survival time of approximately 730 days. We assumed exponential censoring with $\gamma = 1/100$, resulting in heavy censoring with around 90 % of the subjects having censored follow-up times. For log-logistic survival times and exponential censoring, the variance inflation factor is given by

$$\varphi(t) = \frac{\beta}{t^{\beta}} \int_0^t \frac{u^{\beta-1}}{e^{-u\gamma}} du.$$

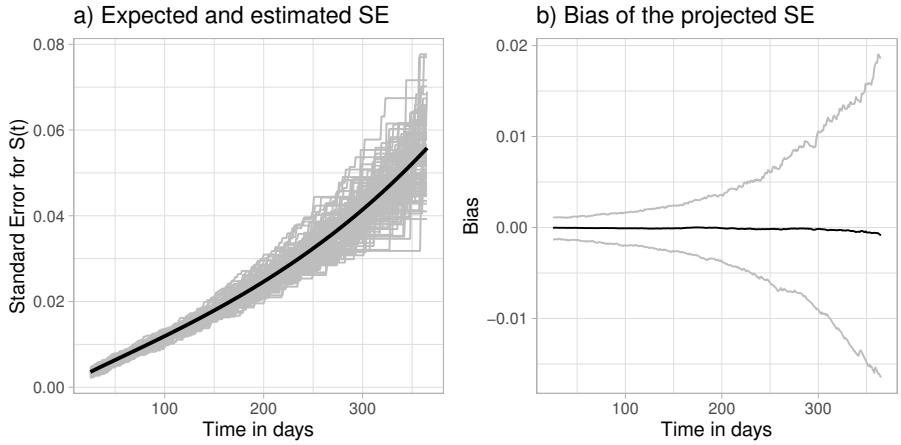

Figure 3 – Results of a simulation study assuming log-logistic survival times with non-monotonic hazard and exponential censoring times depicting (a) the expected (thick black) and estimated (thin grey lines) SE of the survival curve and (b) the bias of the projected SE with 95 % Monte Carlo confidence intervals.

The solution for $\varphi(t)$ results in equations involving the incomplete gamma function, thus we proceeded with numerical integration. Figure 3 presents the results of the simulation study, reiterated with the above described settings 10^4 times. The estimation showed decreasing precision with time, as was expected due to the heavy censoring.

5. Application in research planning

5.1. Confidence intervals estimation

Rajkumar *et al.* (2010) in an open-label randomised controlled trial examined the combination lenalidomide with low or high-dose dexamethasone as therapy for newly diagnosed multiple myeloma. The trial showed that the high-dose group of 221 subjects had higher mortality. This was assessed and illustrated with Kaplan-Meier curves. Likewise, for the sake of visibility the confidence intervals were not included. However, the figure (Fig. 2 in Rajkumar *et al.*, 2010) does present numbers at risk at 6, 12 and 18 months. Using this information and Eq. (5) we could assess standard errors and construct confidence intervals for the survival probabilities. We assumed that the survival times are exponentially distributed. We estimated the hazard as $\lambda = t^{-1}\log(S(t))$, and the assessed hazard at 6, 12 and 18 months to be 0.0139, 0.0135 and 0.0138. Thus we concluded that the survival times were exponentially distributed with a hazard of 0.0137, the average of the 6, 12 and 18 months hazards. Given that $Y(t) = nS(t)G(t)$ we estimated $G(t) = Y(t)(nS(t))^{-1}$. Fitting an exponential distribution to the censoring times suggested accelerated drop-out rate with time. The estimated hazard at 6 months was 0.009, at 12 months 0.05 and at 18 months 0.85. Assuming a Weibull distribution we set up the equation $\log(G(t)) = -(t\theta^{-1})^k$ and estimated the scale ($\theta = 15.05$) and shape ($k = 2.43$) using nonlinear least-squares. We run 10^4 simulations which suggested that the assumed distributions for the survival and censoring times are feasible with a certain bias for censoring at the beginning of the follow-up period (Table 1).

TABLE 1

Survival probabilities ($\hat{S}(t)$) and numbers at risk($\hat{Y}(t)$) reported by Rajkumar et al. (2010) and mean survival probabilities ($S^*(t)$) and numbers at risk($Y^*(t)$) form 10^4 simulations.

	$\hat{S}(t)$	$\hat{Y}(t)$	$S^*(t)$	$Y^*(t)$
6 m	0.92	192	0.92	183
12 m	0.85	103	0.85	105
18 m	0.78	37	0.78	37

TABLE 2

Standard error for the survival probabilities reported by Rajkumar et al. (2010) under the assumption of no censoring (SE(bin)), the variance inflation factor ($\varphi(t)$), the expected variance under the censoring assumption SE(φ)and mean SEs for 10^4 Monte Carlo simulations and the estimated 95 % confidence intervals.

	SE(bin)	$\varphi(t)$	SE(φ)	SE(MC)	95 % CI
6 m	0.0181	1.0328	0.0184	0.0184	0.875; 0.949
12 m	0.0265	1.2100	0.0265	0.0265	0.789; 0.894
18 m	0.0385	1.8159	0.0374	0.0385	0.695; 0.843

Table 2 summarizes the standard errors for 6, 12 and 18 months under the assumption of no censoring. In this case the variance estimation simplifies to binomial variance. The inflation factor increases, as expected, with time almost to factor 2 by 18 months. The SE estimate calculated using Eq. (3) equals the standard error calculated from the Monte Carlo simulations. Using standard confidence interval techniques, it is straightforward to construct confidence intervals.

5.2. *Sample size for open-label single-arm studies*

Once efficacy is proven against placebo or against standard-of-care, new treatments of diseases with high mortality rate can proceed with open-label single-arm studies. Open-label single-arm studies often do not assess sample size with respect to statistical power, but precision, i.e. width of the confidence interval around the point estimate. There are several competing methods for confidence estimation for survival probabilities, however the log transformation of the cumulative hazard is one of the most commonly used (Klein and Moeschberger, 2006). The $100(1-\alpha)\%$ confidence interval for the survival function is given by $\{\hat{S}(t)^{1/\theta}; \hat{S}(t)^{\theta}\}$ where

$$\theta = \exp\left\{\frac{Z_{1-\alpha/2} SE(t) \hat{S}(t)^{-1}}{\ln \hat{S}(t)}\right\}.$$

From Eq. (3) we know that $SE(t) = \sqrt{\varphi(t)\hat{S}(t)\left(1-\hat{S}(t)\right)n^{-1}}$. Using this $SE(t)$ estimator as plug-in and setting the width of a confidence interval to

$$d(t) = \hat{S}(t)^{\theta} - \hat{S}(t)^{1/\theta}, \tag{8}$$

we can use numerical root finding to assess the required sample size to obtain the desired precision around the survival probability estimate.

Idiopathic pulmonary fibrosis (IPF) causes scarring (fibrosis) of the lungs until the lungs cannot take in enough oxygen. The reasons behind IPF are unknown and there is an unmet need of efficacious medication. Current standard-of-care involves treatment with Nintedanib, which showed a numerical reduction (but non-significant) in the risk of all-cause mortality compared to placebo (Richeldi *et al.*, 2018). In Richeldi *et al.* (2018) survival probabilities and numbers at risk at different time points are shown. The survival probability at 52 weeks was 0.66. Together with the survival probabilities, the authors also present numbers at risk at different time points. Using the methodology described in the previous subsection we concluded that survival time is best described by a Weibull distribution with scale parameter of 143.95 and shape parameter of 0.92. Similarly, the censoring can be described by a Weibull distribution. Assuming $S(52w) = 0.66$ and the above listed shape and scale parameter values Eq.(5) gives $\varphi(52w) = 1.606$. For a hypothetical trial with the same efficacy but better safety profile to have a 95 % confidence interval with width of 0.2 at 52 weeks with Eq. (8) and the above listed parameter

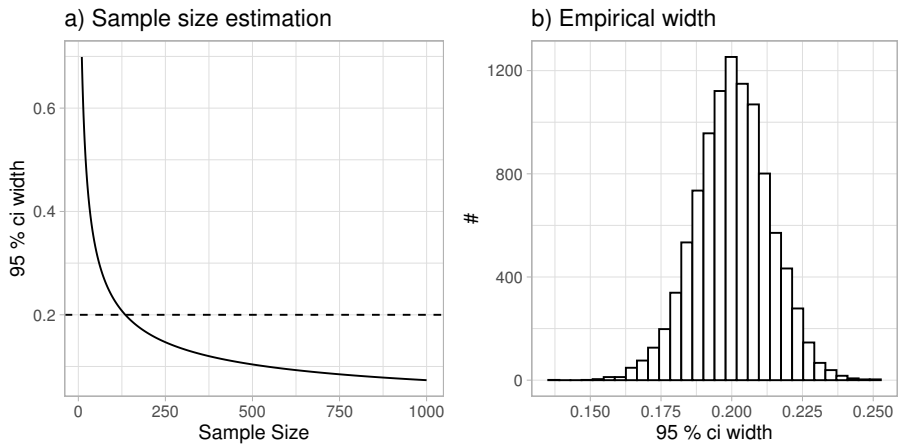

Figure 4 – Sample size estimation for a given confidence interval width under type I censoring (a) and the empirical with of the 95 % confidence interval at 52 weeks based on data from the TOMORROW trial (b).

values we estimated that we need 135 subjects (Figure 4a). Thereafter, we conducted 10^4 simulations, which show that the average width of the confidence intervals at 52 weeks was 0.200 (SD of 0.013) (Figure 4b).

5.3. Estimation under model misspecification

Application of Eq. (5) requires parametric assumptions. While for the endpoint of interest there might be subject specific knowledge or empirical results that favor one or another distribution, generally less information is available for the censoring distribution. Phase II clinical trials with time-to-event endpoints in asthma trials often assume exponential survival times. Empirical data suggest approximately 35 % of subjects will have an event in the placebo arm and up to 15% will be lost for follow-up. In a simulation study with 200 subjects and 12 weeks follow-up we assumed exponential survival times with hazard of 0.0358. We examined the effect of falsely assuming exponential censoring times when the true censoring distribution is Weibull. The scale parameter was set so that $\exp(-(12/\theta)^k) = 0.85$.

Figure 5 summarises the finding of the simulations. As expected, deviations from $k = 1$ when the Weibull distribution simplifies to the exponential distribution induced bias. The bias of the variance inflation factor and the bias of the variance of the Kaplan-Meier survival curve coincide well (Table 3).

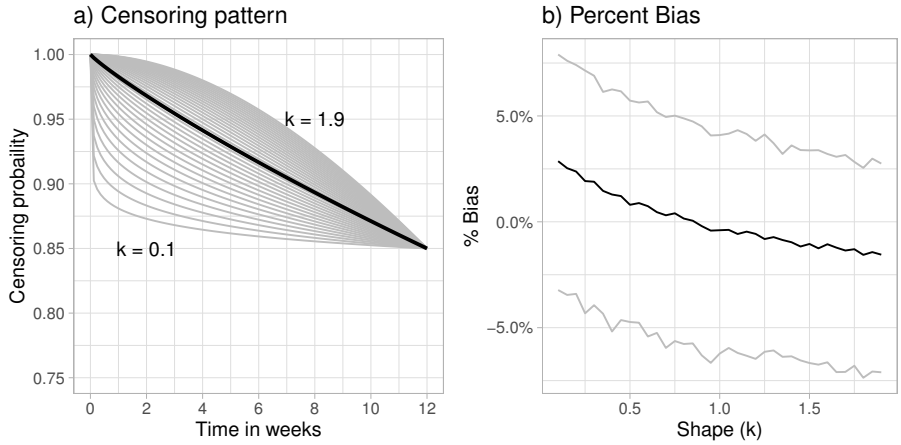

Figure 5 – The effect of model misspecification on the proposed estimator when the underlying distribution pattern is Weibull but calculation assumes exponential censoring distribution (a) and the recorded percent bias of the SE of the survival function at 12 weeks (b).

TABLE 3

The effect of model misspecification for the censoring distribution on the variance inflation factor and the variance of the survival curve at 12 weeks. The assumed censoring distribution is exponential, while the true is Wiebull with shape k and scale θ. The table present the assumed variance inflation ($\varphi(t)$) the $\varphi^*(t)$ and the expected % bias of the inflation factor and the variance of the survival at $t = 12$ weeks.

k	θ	$\varphi(t)$	$\varphi^*(t)$	% Bias $\varphi(t)$	% Bias σ_t^2
0.2	105842.8	1.092	1.148	-4.887	-4.428
0.4	1126.9	1.092	1.128	-3.206	-2.824
0.6	247.9	1.092	1.113	-1.903	-1.285
0.8	116.2	1.092	1.102	-0.858	-0.380
1.0	73.8	1.092	1.092	0.000	0.534
1.2	54.5	1.092	1.084	0.719	0.943
1.4	43.9	1.092	1.078	1.331	1.750
1.6	37.3	1.092	1.072	1.858	2.380
1.8	32.9	1.092	1.067	2.317	2.848

6. Concluding remarks

Herein we have derived a simple to use analytic expression that allows estimation of the variance inflation due to censoring. The advantage of the proposed formula is the relative ease of estimation. For the exponential distribution, that is routinely assumed in most clinical trials, there exists a closed form solution. For other distributions, or combination of distributions, routine programming skills would suffice to obtain estimates.

Irrespective of the distribution assumed we observed that the proposed variance inflation estimator is negatively biased. The projected SE under censoring marginally exceeds the SE of the survival curves estimated by the Greenwood estimator. This is not unexpected as the Kaplan-Meier estimator under right censoring is upward biased. The magnitude of the bias depends on the numbers at risk $Y(t)$, as $exp\{-Y(t)\}$ (Meier, 1975). Klein (1991) and Zhao (1996) assessed the small sample properties of the Greenwood variance estimator and concluded that it is negatively biased. However, the recorded bias was smaller than the competing estimators. Given these aspects we did not attempt to conduct a more in-depth examination of the bias of the proposed estimator. Zhang (1999) and Wang (2016) proposed new unbiased/improved estimators, but these did not gain traction either in the statistical literature or software implementation. While we solely focused on the accuracy of the proposed estimator, there are other aspects that are worth mentioning and perhaps these can be further examined in future publications. For example we did not examine how the proposed estimator would translate to other estimates in survival analysis that are in direct relationship with the Kaplan-Meier estimator (the Nelson -Aalen for the cumulative hazard rate) or of which the Kaplan-Meier estimator is an integral part (Aalen-Johansen estimator for cumulative incidences). Furthermore, herein we did attempt to establish a direct connection between the variance inflation and sample size calculations, a subject that deserves more interest.

As we exemplified, the proposed equation and framework can be used, among other things, to understand the variability around point estimates of survival probability, or to sample size estimations based on confidence interval width. This is particularly interesting, as existing sample size estimation software provide such estimates only under type II censoring, which is useful in engineering but have lite applicability in medical statistics (Leung *et al.*, 1997).

The variance inflation estimator proposed here, is more indicative than informative in nature. This is due to the fact that its estimation requires distributional assumptions. Assumptions that are likely to be violated by real life data. However, this kind of assumptions are routinely made in statistical research design, where we believe this estimator is useful and researchers can have a quick assessment of the excess variability caused by censoring and loss to follow-up.

Appendix

A. Variance estimators under no censoring

If there is no censoring the Greenwood estimator simplifies to the binomial variance

$$\widehat{\text{var}}(S(t)) = \hat{S}(t)\left(1-\hat{S}(t)\right)n^{-1}.$$

To prove this equality we can divide both sides of Eq. (2) with $\hat{S}^2(t)$ giving

$$\frac{1-\hat{S}(t)}{n\hat{S}(t)} = \sum_{T_i \le t} \frac{dN(T_i)}{Y(T)[Y(T_i)-dN(T_i)]}.$$

Under no censorship $Y(0) = n$ and $Y(T_i) = n - \sum dN(T_i)$ and $S(T_i) = Y(T_i)n^{-1}$. The RHS can be rewritten as

$$\sum_{T_i \le t} \frac{dN(T_i)}{Y(T)[Y(T_i)-dN(T_i)]} = \frac{1}{n(n-1)} + \frac{1}{(n-1)(n-2)} + \frac{1}{(n-2)(n-3)} + \ldots$$

thus the increment is

$$\frac{1}{(n-\sum dN(T)+1)(n-\sum dN(T)+1)}.$$

The LHS can be rewritten as

$$\frac{1-\hat{S}(T_i)}{n\hat{S}(T_i)} = \frac{n-Y(T_i)}{nY(T_i)} = \frac{\sum dN(t_i)}{n(n-\sum dN(t_i))},$$

again with the increment

$$\frac{1}{(n-\sum dN(T)+1)(n-\sum dN(T)+1)}.$$

Additionally, we know that $\hat{S}(t)$ and its variance is constant between the observed survival times. This together with the equality of the increments proves that the Greenwood estimator reduces to the binomial variance if there is no censoring in the data. This property of the Greenwood estimator is not shared by other estimators for the variance of survival probabilities. For example Cantor (2001) considered the Peto estimator (Peto *et al.*, 1977) in its modified form due to Slud *et al.* (1984) as $S(t)(1-S(t))Y(t)^{-1}$. As noted above $Y(t) = nS(t)G(t)$ (Andersen *et al.*, 2012) so the Peto estimator is $(1-S(t))(nG(t))^{-1}$. If there is no censoring, i.e. $G(t) \equiv 1$ then the Peto estimator simplifies to $(1-S(t))n^{-1}$, which is not a valid variance estimator for binomial probability.

B. Using the R computing environment as a tool

As we highlighted in subsection 4.1 for exponential survival and censoring times we can obtain closed form solution for the time dependent variance inflation factor $\varphi(t)$ (Eq. 6). However we can use readily available routines of the R computing environment (R Core Team, 2019) to estimate $\varphi(t)$. We assume that both $S(t)$ and $G(t)$ are exponentially distributed with hazard for event λ and hazard for censoring, γ then the variance inflation due censoring can be expressed as

$$\varphi(t) = \frac{e^{-\lambda t}}{1-e^{-\lambda t}} \int_0^t \frac{\lambda}{e^{-\lambda u} e^{-\gamma u}} du.$$

The closed form solution can be easily implemented in any programming language, or we can use numerical integration. We start by defining the survival and censoring hazards, λ and γ, and the follow-up time of interest.

```
lambda <- 1/365  # survival hazard
gamma <- 1/365   # censoring hazard
x <- 50:100      # follow-up times
```

Thereafter we need to define the integrand, the RHS of the equation above

```
integrand <- function(x) {lambda/(exp(-x*lambda)*exp(-x*gamma))}
```

Then we can use numerical integration to estimate $\varphi(t)$

```
phi = vector("numeric", length(x))
for(i in 1:length(x)){
phi[i] <- exp(-x[i]*lambda)/(1-exp(-x[i]*lambda))*
          integrate(integrand, lower = 0, upper = x[i])$value }
```

It is easy to check that the numerical estimates equals the estimates from the closed form solution

```
(lambda/(lambda+gamma))*(exp(x*(lambda+gamma))-1)/(exp(x*lambda)-1)
```

References

O. Aalen, O. Borgan, H. Gjessing (2008). *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag, New York.

M. G. Akritas (2000). *The central limit theorem under censoring*. Bernoulli, 6, no. 6, pp. 1109–1120.

P. K. Andersen, O. Borgan, R. D. Gill, N. Keiding (2012). *Statistical odels Based on Counting Processes*. Springer-Verlag.

R. Brooks (1982). *On the loss of information through censoring*. Biometrika, 69, no. 1, pp. 137–144.

A. Burton, D. G. Altman, P. Royston, R. L. Holder (2006). *The design of simulation studies in medical statistics*. Statistics in Medicine, 25, no. 24, pp. 4279–4292.

A. B. Cantor (2001). *Projecting the standard error of the Kaplan–Meier estimator*. Statistics in Medicine, 20, no. 14, pp. 2091–2097.

M. Greenwood (1926). *A Report on the Natural Duration of Cancer*. HMSO, London.

J. Irwin (1949). *The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice*. Epidemiology & Infection, 47, no. 2, pp. 188–189.

E. L. Kaplan, P. Meier (1958). *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, 53, no. 282, pp. 457–481.

J. P. Klein (1991). *Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators*. Scandinavian Journal of Statistics, pp. 333–340.

J. P. Klein, M. L. Moeschberger (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.

K.-M. Leung, R. M. Elashoff, A. A. Afifi (1997). *Censoring issues in survival analysis*. Annual Review of Public Health, 18, no. 1, pp. 83–104.

P. Meier (1975). *Estimation of a distribution function from incomplete observations*. Journal of Applied Probability, 12, no. S1, pp. 67–87.

P. Meier, T. Karrison, R. Chappell, H. Xie (2004). *The price of Kaplan–Meier*. Journal of the American Statistical Association, 99, no. 467, pp. 890–896.

S. Nadarajah (2003). *Reliability for lifetime distributions*. Mathematical and Computer Modelling, 37, no. 7-8, pp. 683–688.

R. PETO, M. PIKE, P. ARMITAGE, N. E. BRESLOW, D. COX, S. HOWARD, N. MANTEL, K. MCPHERSON, J. PETO, P. SMITH (1977). *Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples*. British Journal of Cancer, 35, no. 1, p. 1.

R CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

S. V. RAJKUMAR, S. JACOBUS, N. S. CALLANDER, R. FONSECA, D. H. VESOLE, M. E. WILLIAMS, R. ABONOUR, D. S. SIEGEL, M. KATZ, P. R. GREIPP (2010). *Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: An open-label randomised controlled trial*. The Lancet Oncology, 11, no. 1, pp. 29–37.

L. RICHELDI, M. KREUTER, M. SELMAN, B. CRESTANI, A.-M. KIRSTEN, W. A. WUYTS, Z. XU, K. BERNOIS, S. STOWASSER, M. QUARESMA (2018). *Long-term treatment of patients with idiopathic pulmonary fibrosis with nintedanib: Results from the tomorrow trial and its open-label extension*. Thorax, 73, no. 6, pp. 581–583.

E. V. SLUD, D. P. BYAR, S. B. GREEN (1984). *A comparison of reflected versus test-based confidence intervals for the median survival time, based on censored data*. Biometrics, pp. 587–600.

W. STUTE (1995a). *The central limit theorem under random censorship*. The Annals of Statistics, 23, no. 2, pp. 422–439.

W. STUTE (1995b). *The statistical analysis of Kaplan-Meier integrals*. Lecture Notes-Monograph Series, 27, pp. 231–254.

W. STUTE (2003). *Kaplan–Meier integrals*. Handbook of Statistics, 23, pp. 87–104.

J. WANG (2016). *Improved versions of Greenwood estimators under Koziol-Green model*. Communications in Statistics-Theory and Methods, 45, no. 1, pp. 142–157.

G. YANG (1977). *Life expectancy under random censorship*. Stochastic Processes and their Applications, 6, no. 1, pp. 33–39.

P. H. ZHANG (1999). *Exact bias and variance of the product limit estimator*. Sankhyā: The Indian Journal of Statistics, Series B, pp. 413–421.

G. ZHAO (1996). *The homogenetic estimate for the variance of survival rate*. Statistics in Medicine, 15, no. 1, pp. 51–60.

G. ZHENG, J. L. GASTWIRTH (2001). *On the Fisher information in randomly censored data*. Statistics & Probability Letters, 52, no. 4, pp. 421–426.

SUMMARY

One of the most obvious features of time-to-event data is the occurrence of censoring. Rarely, if ever, studies are conducted until all participants experience the event of interest. Some participants survive beyond the end of follow-up time, some drop out from the studies for various non-study related reasons. During research planning it is paramount to consider the effect of censoring the follow-up times on the estimates. Herein, we look into the possibility of assessing the loss of information, as measured by the variability of the survival probability estimates under right censoring. We provide the researchers with an easy to use formula to assess the magnitude of variance inflation due to censoring. Additionally, we conducted simulation studies assuming various survival distributions. We conclude that the provided variance inflation estimator can be an accurate practical tool for applied statisticians.

Keywords: Survival; Right-censoring; Standard error; Kaplan-Meier.