The Safe Logrank Test:Error Control under Optional Stopping, Continuation and Prior Misspecification

Peter Grünwald[*] pdg@cwi.nl
Alexander Ly[†] Alexander.Ly@cwi.nl
Muriel Perez-Ortiz Muriel.Perez@cwi.nl
Judith ter Schure Judith.ter.Schure@cwi.nl
Machine Learning Group, CWI Amsterdam[‡]

Abstract

We introduce the safe logrank test, a version of the logrank test that can retain type-I error guarantees under optional stopping and continuation. It allows for effortless combination of data from different trials on different sub-populations while keeping type-I error guarantees and can be extended to define always-valid confidence intervals. Prior knowledge can be accounted for via prior distributions on the hazard ratio in the alternative, but even under 'bad' priors Type I error bounds are guaranteed. The test is an instance of the recently developed martingale tests based on e-values. Initial experiments show that the safe logrank test performs well in terms of the maximal and the expected amount of events needed to obtain a desired power.

1. Introduction

Traditional hypothesis tests and confidence intervals lose their validity under *optional stopping* and *continuation*. Very recently, a new theory of testing and estimation has emerged for which optional stopping and continuation pose no problem at all (Shafer et al., 2011; Howard et al., 2021; Ramdas et al., 2020; Vovk and Wang, 2021; Shafer, 2020; Grünwald et al., 2019). For instantiations of (forerunners of) the ideas developed here within AI and machine learning (where there are obvious applications in e.g. A/B testing), see Balsubramani and Ramdas (2015); Johari et al. (2017). The main ingredients of the new developments are the e-value, a direct alternative to the classical p-value, and the test martingale, a product of conditional e-variables. These are used to create so-called *safe* tests that preserve type-I error control under optional stopping and continuation, and *always-valid* confidence intervals that remain valid irrespective of the stopping time employed. Here we provide a concrete instance of this theory: we develop E-variables and martingales for a safe (under optional stopping) version of the classical logrank test of survival analysis (Mantel, 1966; Peto and Peto, 1972) as well as for regression with Cox's (1972) immortal proportional hazards model — settings in which optional stopping and continuation is highly desirable. At the time of writing, the former of these has already been implemented in an R package (Ly and Turner, 2020). We provide some initial experimental results in Section 4.

[*] Also affiliated with Leiden University, Department of Mathematics.
[†] Also affiliated with University of Amsterdam, Department of Psychology.
[‡] CWI is the National Resarch Institute for Mathematics and Computer Science in the Netherlands.

Logrank tests and proportional hazards are standard tools and assumptions in randomized clinical trials, and are already often combined with group sequential/α-spending approaches. Such approaches allow several interim looks at the data to stop for efficacy or futility. Like ours, they are rooted in early work by H. Robbins and his students (Darling and Robbins, 1967; Lai, 1976), but the details are very different. The advantage of using E-variables instead of α-spending is that the former is still more flexible, and as a consequence easier to use. In particular, with group sequential approaches one has to specify in advance at what points in time one is allowed to do an interim analysis; α-spending is more flexible but still needs a maximum sample size to be set in advance. With E-variables, one can always look and one can always add new data. This becomes especially interesting if one wants to combine the results of several trials in a bottom-up retrospective meta-analysis, where no top-down stopping rule can be enforced : if a randomized clinical trial was reasonably successful but not 100% convincing, then a second randomized trial might be performed *because* of this result— the trials are not independent (Ter Schure and Grünwald, 2019). As a result of the second, a third might be performed, and so on. Even if the alternative hypothesis in all these trials is different (we may have, e.g. different effect sizes in different hospitals), as long as it is of interest to reject a global null (no effect in any trial) we can simply combine all our E-variables of individual trials by multiplication — the resulting test still has a valid type-I error guarantee. Moreover, we can even combine interim results of trials by multiplication while these trials are still ongoing — going significantly beyond the realm of traditional α-spending approaches. We also show how E-variables can be combined with Bayesian priors, leading to nonasymptotic frequentist type-I error control even if these priors are wildly misspecified (i.e. they predict very different data from the data we actually observe). Our approach is sequential in nature, but significantly more flexible than earlier sequential approaches such as Jones and Whitehead (1979) and Sellke and Siegmund (1983). This, and many other details for which there is no space in this extended abstract are treated in the extended arXiv version of this paper Grunwald et al. (2020) (ARX from now on). We refer to Grünwald et al. (2019) (GHK from now on) for an extensive general introduction to E-variables including their relation to likelihood ratios (when both the null hypothesis $\mathcal{H}_0$ and the alternative $\mathcal{H}_1$ are simple (singleton), then the best E-variable coincides with the likelihood ratio); Bayes factor hypothesis testing (E-variables are often, but not always, Bayes factors; and Bayes factors are often, but not always E-variables) and their enlightening *betting* interpretation (indeed, e-values are also known under the name *betting scores* Shafer (2020)). The general story that emerges from papers such as Shafer's as well as GHK and Ramdas et al. (2020) is that E-variables and test martingales are the 'right' generalization of likelihood ratios to the case that both $\mathcal{H}_0$ and $\mathcal{H}_1$ can be composite, providing an intuitive notion of evidence.

Contributions We show that Cox' partial likelihood underlying his proportional hazards model can be used to define E-variables and test martingales. In this extended abstract, we only show this in a simplified, discrete time setup for the case without covariates, leading to a 'safe' (for optional stopping) logrank test. In the full version of this paper (ARX) we extend this derivation to the case with unordered simultaneous events (ties), continuous time, 'always-valid' confidence sequences and covariates (Cox regression).

Contents We first provide a short introduction to E-variables and test martingales. In Section 3 we develop E-variables for proportional hazards without covariates, based on Cox' partial likelihood. Section 4 provides some simulations showing the feasibility of our approach in practice, if a minimum statistical power is required.

2. E-Variables and Test Martingales

Before trying to digest the following definition, it may help to consider a simplified setting (different from the survival analysis setting below) in which the $Y_{\langle i \rangle}$ are i.i.d. $\sim P$ with density p under $\mathcal{H}_0 = \{P\}$ and i.i.d. $\sim Q$ with density q under $\mathcal{H}_1 = \{Q\}$. Then the likelihood ratio for the i-th data point $S_{\langle i \rangle} := q(Y_{\langle i \rangle})/p(Y_{\langle i \rangle})$ is an e-variable according to the definition below since $\mathbf{E}_P[S_{\langle i \rangle}] = \int (q(y_i)/p(y_i))p(y_i)dy_i = \int q(y_i)dy_i = 1$, and since data are i.i.d. under the null, $S_{\langle i \rangle}$ is also an e-variable conditional on $Y_{\langle 1 \rangle}, \ldots, Y_{\rangle i-1 \rangle}$. S^i, called test martingale below, is then simply the likelihood ratio $q(Y_{\langle 1 \rangle}, \ldots, Y_{\langle i \rangle})/p(Y_{\langle 1 \rangle}, \ldots, Y_{\langle i \rangle})$.

Definition 1 *Let $\{Y_{\langle i \rangle}\}_{i \in \mathbb{N}_0}$ represent a discrete-time random process and let $\mathcal{H}_0$, the null hypothesis, be a collection of distributions for this process. Fix $i > 0$ and let $S_{\langle i \rangle}$ be a non-negative random variable that is determined by $(Y_{\langle 0 \rangle}, \ldots, Y_{\langle i \rangle})$, i.e. there exists a function f such that $S_{\langle i \rangle} = f(Y_{\langle 0 \rangle}, \ldots, Y_{\langle i \rangle})$. We say that $S_{\langle i \rangle}$ is an E-variable conditionally on $Y_{\langle 0 \rangle}, \ldots, Y_{\langle i \rangle}$ if for all $P \in \mathcal{H}_0$,*

$$\mathbf{E}_P\left[S_{\langle i \rangle} \mid Y_{\langle 0 \rangle}, \ldots, Y_{\langle i-1 \rangle}\right] \leq 1. \tag{1}$$

If (1) holds with equality, we call the E-variable sharp. *If, for each i, $S_{\langle i \rangle}$ is an E-variable conditional on $Y_{\langle 0 \rangle}, \ldots, Y_{\langle i-1 \rangle}$, then we say that the product process $\{S^i\}_{i \in \mathbb{N}}$ with $S^i = \prod_{k=1}^{i} S_{\langle k \rangle}$ is a test supermartingale relative to $\{Y_{\langle i \rangle}\}_{i \in \mathbb{N}_0}$ and the given $\mathcal{H}_0$. If all constituent E-variables are sharp, we call the process a test martingale.*

It is easy to see (Shafer et al., 2011) that a test (super-) martingale is, in more standard terminology, a nonnegative (super-) martingale relative to the filtration induced by $\{Y_{\langle i \rangle}\}_{i \in \mathbb{N}_0}$, with starting value 1.

Safety The interest in E-variables and test martingales derives from the fact that we have type-I error control irrespective of the stopping rule used: for any test (super-) martingale $\{S^i\}_{i \in \mathbb{N}}$ relative to $\{Y_{\langle i \rangle}\}_{i \in \mathbb{N}_0}$ and $\mathcal{H}_0$, Ville's inequality (Shafer, 2020) tells us that, for all $0 < \alpha \leq 1$, $P \in \mathcal{H}_0$,

$$P(\text{there exists } i \text{ such that } S^i \geq 1/\alpha) \leq \alpha.$$

Thus, if we measure evidence against the null hypothesis after observing i data units by S^i, and we reject the null hypothesis if $S^i \geq 1/\alpha$, then our type-I error will be bounded by α, no matter what stopping rule we used for determining i. We thus have type-I error control even if we use the most aggressive stopping rule compatible with this scenario, where we stop at the first i at which $S^i \geq 1/\alpha$ (or we run out of data, or money to generate new data). We also have type-I error control if the actual stopping rule is unknown to us, or determined by external factors independent of the data $Y_{\langle i \rangle}$ — as long as the decision whether to stop depends only on past data, and not on the future (the potential to take into account external factors is not directly visible from Ville's inequality as stated here; it is formalized by GHK19).

We will call any test based on $\{S^i\}_{i \in \mathbb{N}}$ and a (potentially unknown) stopping time τ that, after stopping, rejects iff $S^\tau \geq 1/\alpha$ a *level α-test that is safe under optional stopping*, or simply a *safe test*. Note that in our simple i.i.d. example above with $cH_0 = \{P\}$ and $\mathcal{H}_1 = \{Q\}$, the most power Neyman-Pearson test at level α is also a likelihood ratio test, but with threshold *that also depends on sample size n* — for us, the threshold is $1/\alpha$ irrespective of n, and this is the key to enabling optional stopping. Importantly, we can also deal with *optional continuation*: we can combine E-variables from different trials that share a common null (but may be defined relative to a different alternative) by multiplication, and still retain type-I error control (we give examples in ARX). If we used p-values instead would have to resort to e.g. Fisher's method, which, in contrast to multiplication of e-values, is invalid if there is a dependency between the (decision to perform) tests.

Optimality Just like for p-values, the definition of E-variables only requires specification of $\mathcal{H}_0$, not of an alternative hypothesis $\mathcal{H}_1$. $\mathcal{H}_1$ comes into play once we distinguish between 'good' and 'bad' E-variables: E-variables have been designed to remain small, with high probability, under the null $\mathcal{H}_0$. But if $\mathcal{H}_1$ rather than $\mathcal{H}_0$ is true, then 'good' E-variables should produce evidence (grow — because the larger the E-variable, the closer we are to rejecting the null) against $\mathcal{H}_0$ as fast as possible. First consider a simple (singleton) $\mathcal{H}_1 = \{P\}$. If data comes from P, then the optimality of conditional E-variable $S_{\langle i \rangle}$ is measured in terms of $\mathbf{E}_P[\log S_{\langle i \rangle} \mid Y_{\langle 0 \rangle}, \ldots, Y_{\langle i-1 \rangle}]$. The E-variable which maximizes this quantity among all E-variables is called *Growth Rate Optimal in the Worst case*, GROW. There are various reasons why one should take a logarithm here — see GHK and Shafer (2020) for details. We explore one in detail in ARX: by *Wald's identity*, among all E-variables, the GROW minimizes the expected number of data points needed before the null can be rejected. Thus, finding a sequence of GROW E-variables is quite analogous to finding the test that maximizes power — in Section 4 we provide some simulations to relate power to GROW. Note that we cannot directly use power itself in designing tests, since the notion of power requires a fixed sampling plan, which by design we do not have. In case $\mathcal{H}_1$ is composite, we extend the notion of GROW to yield optimal growth in the worst case: the GROW E-variable for outcome i conditional on $Y_{\langle 0 \rangle}, \ldots, Y_{\langle i-1 \rangle}$, if it exists, is the E-variable S that achieves

$$\max_{S} \min_{P \in \mathcal{H}_1} \mathbf{E}_P[\log S_{\langle i \rangle} \mid Y_{\langle 0 \rangle}, \ldots, Y_{\langle i-1 \rangle}], \tag{2}$$

the maximum being over all E-variables conditional on $Y_{\langle 0 \rangle}, \ldots, Y_{\langle i-1 \rangle}$.

3. Safe Logrank Tests

Preliminaries Throughout the text we abbreviate $\{1, \ldots, n\}$ to $[n]$. We assume that n participants are included in a trial, with groups 1 (treatment) and 0 (control). We let $\boldsymbol{g} = (g_1, \ldots, g_n)$ be the binary vector indicating for each participant what group they were put into. In the general continuous time set-up, random variable T_j denotes the time at which the event happens for participant j. All our results continue to hold under noninformative right censoring. For simplicity, we will omit it from our treatment here, but we take it into account in ARX.

We let $Y_j(t) = \mathbf{1}_{T_j \geq t}$, be the 'at risk' process for the j-th participant, and let Y^g be the number of participants at risk in the group $g \in \{0,1\}$ at time t, that is, $Y^g(t) = \sum_{j:\boldsymbol{g}_j = g} Y_j(t)$. We define $\boldsymbol{Y}(t) = (Y_1(t), \ldots, Y_n(t))$ to be the n-dimensional indicator vector that indicates for each participant j whether the participant is still at risk at time t. We set $N^g[t', t] = Y^g(t') - Y^g(t)$ to be the number of events that happened in group g inbetween time t' and time t. We assume that a time increment of 1 represents a natural 'unit time' for example an hour, a day, or a week.

3.1. The Simplified Process in discrete time

In any particular realization of the setting above, we will have a sequence of event times $t\langle 1\rangle < t\langle 2\rangle < t\langle 3\rangle < \ldots$ such that for all i, at time $t\langle i\rangle$, one or more events happen, and inbetween $t\langle i\rangle$ and $t\langle i+1\rangle$, no events happen. We extend the notation to $N^g_{\langle i\rangle}$ to denote the number of events happening in group g at the i $^{\text{th}}$ event time and $\boldsymbol{Y}_{\langle i\rangle} = (Y_{1,\langle i\rangle}, \ldots, Y_{n,\langle i\rangle})$ with $Y_{j,\langle i\rangle} = 1$ if $T_j \geq t\langle i\rangle$. Thus $Y_{j,\langle 0\rangle} = 1$ for all $j \in [n]$, $Y_{j,\langle 1\rangle} = 1$ for all $j \in [n]$ except one, and so on, assuming no censoring: at the time of the first event, everyone is at risk; at the time of the second event, everyone is at risk except the participant that had the first event, etc. Again, $\boldsymbol{Y}_{\langle i\rangle}$ is the n-dimensional vector that indicates for each participant j whether they are still at risk, but now at the time that the i^{th} event happens. Let $Y^g_{\langle i\rangle}$ be the number of participants at risk in the group $g \in \{0,1\}$ at the time of the i^{th} event, that is, $Y^g_{\langle i\rangle} = \sum_{j:g_j=g} Y_{j,\langle i\rangle}$.

Our method is best explained by first assuming that at each time $t\langle i\rangle$, exactly one event happens so $N^0_{\langle i\rangle} + N^1_{\langle i\rangle} = 1$, allowing us to abstract away from 'absolute' time scales. We can then define the *simplified process* $\boldsymbol{Y}_{\langle 0\rangle}, \boldsymbol{Y}_{\langle 1\rangle}, \ldots$ with each $\boldsymbol{Y}_{\langle i\rangle}$ taking values in $\{0,1\}^n$ — note that this process is defined relative to a discrete sample space $[n]^\infty$ in which there is no notion of continuous time. For given group assignment $\boldsymbol{g}$ and each $\theta > 0$ we define a distribution P_θ underlying this process such that:

1. $\boldsymbol{Y}_{\langle 0\rangle} = (1, 1, \ldots, 1)$, P_θ-a.s.
2. For each $i \leq n$, there is a single participant $j^\circ \in [n]$ that experiences an event, i.e. we have $Y_{j^\circ,\langle i\rangle} = 0, Y_{j^\circ,\langle i-1\rangle} = 1$, and for all $j \in [n]$ with $j \neq j^\circ$, $Y_{j^\circ,\langle i\rangle} = Y_{j^\circ,\langle i-1\rangle}$. We let $J_{\langle i\rangle} = j^\circ$ be the RV denoting this participant.
3. We set for j° with $g_{j^\circ} = 1$: $P_\theta(J_{\langle i\rangle} = j^\circ \mid Y_{j^\circ,\langle i-1\rangle} = 1) := \frac{\theta}{Y^0_{\langle i-1\rangle} + Y^1_{\langle i-1\rangle}\theta}$ and for j° with $g_{j^\circ} = 0$: $P_\theta(J_{\langle i\rangle} = j^\circ \mid Y_{j^\circ,\langle i-1\rangle} = 1) = \frac{1}{Y^0_{\langle i-1\rangle} + Y^1_{\langle i-1\rangle} \cdot \theta}$.

These requirements uniquely specify P_θ. In ARX we motivate the definition above as giving essentially the correct conditional distribution of $J_{\langle i\rangle}$ under a proportional hazards assumption with hazard ratio θ. We define q_θ to be the conditional probability mass function of the event that the i-th event takes place in group g. That is:

$$q_\theta(g \mid (y^0, y^1)) := P_\theta(N^g_{\langle i\rangle} = 1 \mid Y^0_{\langle i-1\rangle} = y^0, Y^1_{\langle i-1\rangle} = y^1)$$

By the above,

$$q_\theta(1 \mid (y^0, y^1)) = 1 - q_\theta(0 \mid (y^0, y^1)) = \frac{y^1\theta}{y^0 + y^1\theta} \tag{3}$$

is the probability mass function of a Bernoulli $y^1\theta/(y^0+y^1\theta)$-distribution; note also that, for any vector $\boldsymbol{y}$ that is compatible with the given y^0, y^1 and $\boldsymbol{g}$, we have $q_\theta(1 \mid (y^0,y^1)) = P_\theta(N^g_{\langle i\rangle} = g \mid \boldsymbol{Y}_{\langle i-1\rangle} = \boldsymbol{y})$: the probability of an event in group g only depends on the counts in both groups. For given $\theta_0,\theta_1>0$, let $M_{\theta_1,\theta_0}\langle 0\rangle = 1$ and

$$M_{\theta_1,\theta_0}\langle i\rangle = \frac{q_{\theta_1}(N^1_{\langle i\rangle} \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})}{q_{\theta_0}(N^1_{\langle i\rangle} \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})}. \tag{4}$$

By writing out the expectation, we see that

$$\mathbf{E}_{P_{\theta_0}}\left[M_{\theta_1,\theta_0}\langle i\rangle \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})\right] = \sum_{g\in\{0,1\}} q_{\theta_1}(g \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle}) = 1. \tag{5}$$

This standard argument immediately shows that, under P_{θ_0}, for all i, all $\theta_1>0$, $M_{\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on $\boldsymbol{Y}_{\langle 0\rangle},\ldots,\boldsymbol{Y}_{\langle i-1\rangle}$, and

$$M^{\langle i\rangle}_{\theta_1,\theta_0} := \prod_{j=1}^{i} M_{\theta_1,\theta_0}\langle i\rangle \tag{6}$$

is a test martingale under P_{θ_0} relative to process $(\boldsymbol{Y})_{i\in\mathbb{N}_0}$. Thus, by Ville's inequality, we have the highly desired:

$$\tilde{P}_{\theta_0}\left(\text{there exists } i \text{ with } M^{\langle i\rangle}_{\theta_1,\theta_0} \geq \alpha^{-1}\right) \leq \alpha. \tag{7}$$

To give a first idea of its use in testing and estimation, we give several examples below, simply acting as if M_{θ_1,θ_0} would also be a test martingale under the unknown true distribution, even though the latter is defined on continuous time. We show that the latter is justified in ARX.

Some of the examples require a generalization of M_{θ_1,θ_0} in which q_{θ_1} in (4) is replaced by another conditional probability mass function $r_i(x \mid y^1,y^0)$ on $x\in\{0,1\}$, allowed to depend on i. For any given sequence of such conditional probability mass functions, $\{r_i\}_{i\in\mathbb{N}}$, we extend definition (4) to $M_{r,\theta_0}\langle 0\rangle = 1$ and

$$M_{r,\theta_0}\langle i\rangle = \frac{r_i(N^1_{\langle i\rangle} \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})}{q_{\theta_0}(N^1_{\langle i\rangle} \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})}. \tag{8}$$

For any choice of the r_i, (5) clearly still holds for the resulting M_{r,θ_0}, making $M_{r,\theta_0}\langle i\rangle$ a conditional E-variable and its product a martingale; and then Ville's inequality (7) must also still hold.

Example 1 [GROW alternative] *The simplest possible scenario is that of a one-sided test between 'no effect' ($\theta_0=1$) and a one-sided alternative hypothesis $\mathcal{H}_1=\{P_{\theta_1}:\theta_1\in\Theta_1\}$ represented by For example, if 'event' means that the participant gets ill, then we would hope that under the treatment, $\underline{\theta}_1$ would be a value smaller than 1 and we would have $\Theta_1=\{\theta: 0<\theta\leq\underline{\theta}_1\}$. If 'event' means 'cured' then we would typically set $\Theta_1=\{\theta:\bar{\theta}_1\leq\theta<\infty\}$ for some $\bar{\theta}_1>1$. We will take the left-sided alternative with $\underline{\theta}<1$ as a running example, but* *everything we say in the remainder of this paper also holds for the right-sided alternative. In the left-sided setting, setting, $M_{\theta_1,1}\langle i\rangle$ is a conditional E-variable for arbitrary $\theta_1 > 0$. More generally, $M_{r,1}\langle i\rangle$ is a conditional E-variable for arbitrary conditional mass functions r_i. Still, the so-called GROW (growth-optimal in worst-case) E-variable as in (2) is given by taking $M_{\underline{\theta}_1,1}$, i.e. it takes the $\theta \in \Theta_0$ closest to θ_0. That is,*

$$\max_{\theta>0}\min_{\theta_1\in\Theta_1} \mathbf{E}_{P_{\theta_1}}[\log M_{\theta,\theta_0\langle i\rangle} \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle}] =$$
$$\max_{\{r_i\}}\min_{\theta_1\in\Theta_1} \mathbf{E}_{P_{\theta_1}}[\log M_{r,\theta_0\langle i\rangle} \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle}]$$

is achieved by setting $\theta = \underline{\theta}$, no matter the values taken by $Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle}$. Here the second maximum is over all sequences of conditional distributions r_i as used in (8). Thus, among all E-variables of the general form $M_{r,1}\langle i\rangle$ there are very strong reasons why setting $r_i = q_{\underline{\theta}}$ is the best one can do. Nevertheless, if one does not restrict oneself to E-variables of the form M_{θ_1,θ_0}, but uses the more general M_{r,θ_0} instead, one may sometimes opt for another 'almost' GROW choice, as elaborated in the next example.

Now suppose we want to do a two-sided test, with alternative hypothesis $\{P_{\theta_1} : \theta_1 \leq \underline{\theta}_1 \vee \theta_1 \geq \bar{\theta}_1\}$ with $\bar{\theta}_1 > 1$. For this case, one can create a new 'combined GROW' E-variable

$$M'\langle i\rangle := \frac{1}{2}\left(M_{\underline{\theta}_1,\theta_0}\langle i\rangle + M_{\bar{\theta}_1,\theta_0}\langle i\rangle\right),$$

which is a conditional E-variable since $\mathbf{E}_{P_{\theta_0}}\left[M_{\theta_1,\theta_0}\langle i\rangle \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})\right] = 1$ (see GHK).

Example 2 [Tests based on Bayesian priors with Frequentist Type-I Error Guarantees] *Now suppose we do not have a very clear idea of which parameter $\theta_1 \in \Theta_1$ to pick; we might thus want to put a prior probability distribution on Θ_1. To accommodate for this we extend our definition (3) to*

$$q_W(1 \mid y^0, y^1) = \int_\theta q_\theta(1 \mid y^0, y^1) dW(\theta)$$

for probability distributions W on $\mathbb{R}$. No matter what W we pick, the resulting $M_{W,\theta_0}\langle i\rangle = q_W(1 \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})/q_{\theta_0}(1 \mid Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle})$ is still an E-variable, as argued above Example 1. If data come from some distribution with parameter $\theta_1 \in \Theta_1$, then M_{W,θ_0} will not be GROW unless W puts all of its mass on θ_1; nevertheless, M_{W,θ_0} can come quite close to the optimal for a whole range of θ_1 and may thus sometimes be preferable over choosing $M_{\underline{\theta}_1,\theta_0}$ — we provide simulations to this end in ARX.

Starting with a prior distribution W with density w, we can use Bayes theorem to derive a posterior distribution $w_i(\theta)$ on Θ_1

$$w_i(\theta) := w(\theta \mid Y^1_{\langle 0\rangle}, Y^0_{\langle 0\rangle}, \ldots, Y^1_{\langle i-1\rangle}, Y^0_{\langle i-1\rangle}) = \frac{\prod_{k=1}^{i-1} q_\theta(N^1_{\langle k\rangle} \mid Y^1_{\langle k-1\rangle}, Y^0_{\langle k-1\rangle}) w(\theta)}{\int_\theta \prod_{k=1}^{i-1} q_\theta(N^1_{\langle k\rangle} \mid Y^1_{\langle k-1\rangle}, Y^0_{\langle k-1\rangle}) w(\theta) d\theta}.$$

We thus get that $q_{W_{i+1}}(1 \mid Y^1_{\langle i\rangle}, Y^0_{\langle i\rangle})$ is equal to

$$\int_\theta q_\theta(1 \mid Y^1_{\langle i\rangle}, Y^0_{\langle i\rangle}) w_{i+1}(\theta) d\theta = \int_\theta q_\theta(1 \mid Y^1_{\langle i\rangle}, Y^0_{\langle i\rangle}) \cdot \frac{\prod_{k=1}^{i} q_\theta(N^1_{\langle k\rangle} \mid Y^1_{\langle k-1\rangle}, Y^0_{\langle k-1\rangle}) w(\theta)}{\int_\theta \prod_{k=1}^{i} q_\theta(N^1_{\langle k\rangle} \mid Y^1_{\langle k-1\rangle}, Y^0_{\langle k-1\rangle}) w(\theta) d\theta} \, d\theta$$

and, by telescoping, $M^{\langle i \rangle}_{W_1,\theta_0}$ is seen to be equal to

$$\frac{\int_\theta \prod_{k=1}^{i} q_\theta(N^1_{\langle k \rangle} \mid Y^1_{\langle k-1 \rangle}, Y^0_{\langle k-1 \rangle}) w(\theta) d\theta}{\prod_{k=1}^{i} q_\theta(N^1_{\langle k \rangle} \mid Y^1_{\langle k-1 \rangle}, Y^0_{\langle k-1 \rangle})}$$

This approach resembles a Bayes-factor in the sense that it involves priors and subjective choices. It is not *Bayesian though in the important sense that our frequentist type-I error guarantee continues to hold, irrespective of the prior we choose. Rather, there is an element of what has been called* luckiness *in the machine learning theory literature (Grünwald and Mehta, 2019): if the prior W turns out 'correct', in the weak sense that the E-variable grows about as fast as we would expect in expectation over the prior, then we get a strongly growing E-variable and will need few events before we can reject the null. If the prior is 'wrong', we need a larger sample. Yet, the type-I error guarantee always holds, also in this 'misspecified' case.*

Now, suppose we do have a minimum clinically relevant $\underline{\theta}_1$ in mind, but we want to exploit favorable situations in which the effect size is even larger than indicated by θ_1, i.e. the 'true' $\underline{\theta}_1$ satisfies $|\underline{\theta} - \theta_0| \geq |\underline{\theta}_1 - \theta_0|$ — these are 'favorable' because we can expect the data to contain more evidence against the null. We may then choose to take a prior that is (strongly) peaked at $\underline{\theta}_1$, but stil places some mass on more extreme values of θ_1.

4. Some Simulations

In this section we report the results of simulating data from the simplified process introduced in Section 3. In ARX we show that this is equivalent to simulating data from a continuous time counting process under proportional hazards. Recall that if we are testing some fixed θ_1 with $\theta_1 \leq 1$ against $\theta_0 = 1$, and we have witnessed k events, the odds of next event happening in group 1 are $\theta_1 Y^1_{\langle k \rangle} : Y^0_{\langle k \rangle}$ under the alternative hypothesis. Thus, simulating in which group the next event happens only takes a (biased) coin flip.

We limit our attention in this section to the aforementioned one-sided testing scenario θ_1 (for some $\theta_1 \in (0,1)$) vs. $\theta_0 = 1$, and we fix our desired level to $\alpha = 0.05$. We consider the stopping rule $\tau_{\theta_1} = \inf\{i : M^{\langle i \rangle}_{\theta_1,1} \geq 1/\alpha\}$, that is, we stop as soon as our test martingale crosses the threshold $1/\alpha$ (aggresive optional stopping). By our previous discussion, we have a type-I error guarantee for this and any other stopping rule. However $\tau_{\theta_1,1}$ may often be too large: it may not be feasible financially or time-wise to wait either until the stopping moment or until we run out of patients to reach a decision. Thus it seems reasonable to determine a number of events $i_{\max}$ after which we stop anyway, and decide to accept the null, even if our test martingale $M^{\langle i \rangle}_{\theta_1,1}$ may have crossed the threshold $1/\alpha$, had we continued the study. We would like to control the probability β of this type-II error, induced by stopping at $\tau_{\theta_1} \wedge i_{\max}$ instead of stopping at τ_{θ_1}. A moment's thought shows that we look for the smallest $i_{\max}$ such that $P_{\theta_1}(\tau_{\theta_1} \geq i_{\max}) \leq 1 - \beta$ for a target power $1 - \beta$, which we fix to 0.8. Of course $i_{\max}$ is just the $(1 - \beta)$-quantile of τ_{θ_1}, and can be determined by repeated simulation in a straightforward manner. We simulate a number of realizations i_{sim} of τ_{θ_1} and use the $(1 - \beta)$-quantile of the observed empirical distribution of τ_{θ_1}. For each configuration $\theta_1, Y^1_{\langle 0 \rangle}, Y^0_{\langle 0 \rangle}$ that we considered, we performed $m_{\text{sim}} = 10000$ simulations and

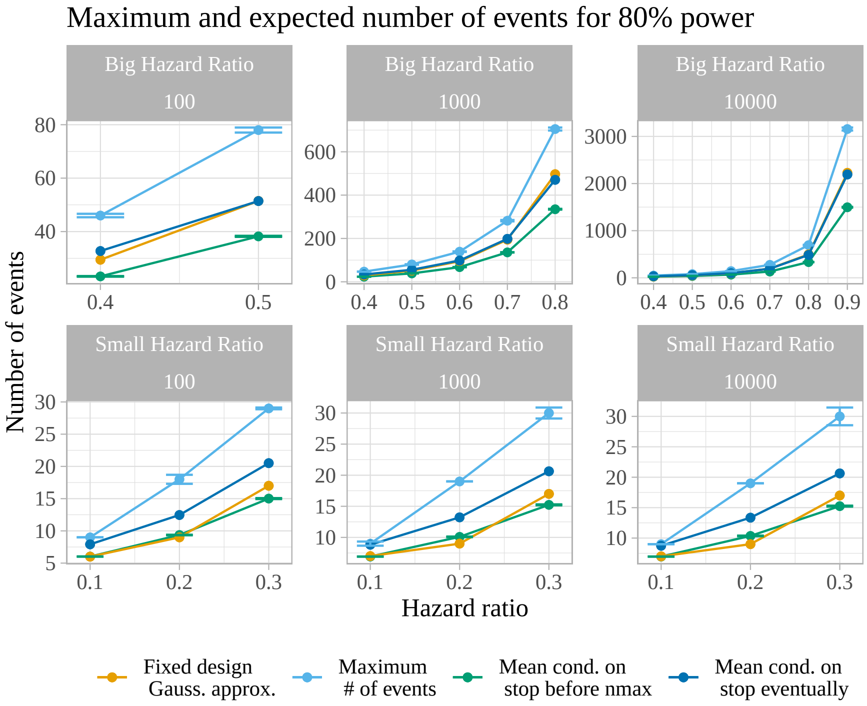

Figure 1: Each panel corresponds to a different starting number of subjects in both groups, e.g. in the left-most panel both groups are of size 100.

assessed the uncertainty in the estimate of $i_{\max}$ by estimating its standard deviation using 1000 bootstrap rounds on the empirical distribution of τ_{θ_1}.

The number of events $i_{\max}$ is the maximum that one may see under the alternative hypothesis at a fixed power $1-\beta$. In this sense, it is the number of events that we will witness in the worst-case. However, we will typically reach a decision sooner. In Figure 1 we show the expected value of the random number of events before we stop, $\tau_{\theta_1} \wedge i_{\max}$, under the null hypothesis (dark blue curve).

For comparison, we also show the number of events that one would need under the Gaussian non-sequential approximation of Schoenfeld (1981) to achieve a power of 0.8 — i.e. one treats the log-rank statistic as if it were normally distributed, and, for fixed number of events, one rejects the null using a z-test, i.e. if the log rank statistic is larger than $z_{0.05} = 1.645$. One then calculates power under the assumption that the log rank statistic also has a normal distribution under the alternative. This is a standard classical approach; see ARX for further details. We see that $i_{\max}$ is significantly larger than the Schoenfeld's predicted number of events, but the expected value of $\tau_{\theta_1} \wedge i_{\max}$, which is the number we will need on average if we plan on stopping at $i_{\max}$ at the latest (dark blue line), is of comparable size.

As we noted earlier, it may happen that data come from a distribution with a more extreme hazard ratio than we anticipated. Then the best choice (the one that leads to the smallest stopping time $\tau_{\theta'_1}$) is to use for our test martingale $M_{\theta'_1,1}$ the value of θ'_1 that actually generates the data. This value is of course unknown in all practical situations. In ARX we provide additional experiments where, to profit from the 'lucky' situation in which the alternative is more extreme than we anticipated, we put a prior W on θ'_1 and use $M_{W,1}$. We find that in this case, we need only slightly more data before we can stop than if we had used $M_{\theta'_1}$ for the θ' that actually generated the data, irrespective of what this θ'_1 is — thus showing that we can achieve the adaptivity of a Bayesian approach while keeping frequentist Type-I error control at the same time.

5. Acknowledgements

This research was supported by the Dutch Research Council (NWO) via research project 617.001.651, *Safe Bayesian Learning.*

References

Akshay Balsubramani and Aaditya Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. *UAI (Uncertainty in Artificial Intelligence) 2015*, 2015.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34(2):187–220, 1972.

D.A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66, 1967.

P. Grünwald and N. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the Thirtieth Conference on Algorithmic Learning Theory (ALT) 2019*, 2019.

P. Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing, 2019. arXiv preprint arXiv:1906.07801.

P Grunwald, Alexander Ly, Muriel F Pérez-Ortiz, and Judith ter Schure. The safe log rank test: Error control under optional stopping, continuation and prior misspecification. *arXiv preprint arXiv:2011.06931*, 2020.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *Annals of Statistics*, 2021. To Appear.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

David Jones and John Whitehead. Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika*, 66(1):105–113, 1979.

Tze Leung Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976.

A. Ly and R. Turner. R-package `safestats`, 2020. install in R by `devtools::install_github( "AlexanderLyNL/safestats", ref = "logrank", build_vignettes = TRUE)`.

Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, 50:163–170, 1966.

Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2):185–198, 1972.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

David Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316–319, 1981.

Judith Ter Schure and Peter Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8, 2019.

Thomas Sellke and David Siegmund. Sequential analysis of the proportional hazards model. *Biometrika*, 70(2):315–326, 1983. Publisher: Oxford University Press.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, 2020. To Appear.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 2021. To Appear.