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

Peter D. Grünwald*† Alexander Ly*‡ Muriel F. Pérez-Ortiz*
Judith ter Schure*

November 16, 2020

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 while keeping type-I error guarantees, and can be extended to define always-valid confidence intervals. The test is an instance of the recently developed martingale tests based on e-values. We demonstrate the validity of the underlying nonnegative martingale and show how to extend it to sequences of events with ties and to Cox' proportional hazards regression. 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). The main ingredients 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. Pace and Salvan (2019) argue that even without optional stopping, always-valid confidence intervals may be preferable over standard ones.

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) proportional hazards model. 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 5.

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

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

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, and thus to some extent related to earlier sequential analyses such as Jones and Whitehead (1979) and Sellke and Siegmund (1983) — although such analyses typically rely on using a precise stopping rule, whereas we allow arbitrary ones, and cannot be easily combined with prior distributions, whereas ours can. One thing that we currently cannot provide for, is dealing with staggered entries of single participants. All these features of our approach are highlighted in Example 1–5 in Section 2, and we discuss staggered entries further in Section 7.

We refer to Grünwald et al. (2019) (GHK from now on) for an extensive 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 — of the many existing generalizations of likelihood to such cases, those that are not E-variables will automatically show problematic behaviour in terms of error control, and those that are can be combined freely over experiments while retaining type-I error control, thereby providing an intuitive notion of evidence.

Contributions We show that Cox' partial likelihood underlying his proportional hazards model defines E-variables and test martingales. We first, in Section 2.2, show this (a) for the case without covariates and without unordered simultaneous events (ties) , leading to a 'safe' (for optional stopping) logrank test. We then, (b), in Section 2.3, extend this to the case with ties, and (c), in Section 3, to the case with covariates. To keep the story simple, we consider time discretized to arbitrary but finite precision, but for completeness, in Section 6 we give a completely general proof, with continuous time, for case (a). Case (a) and (b) vary on existing results and may not be so surprising to readers familiar with martingale theory — though note that we work with nonnegative martingales, which is different from most traditional uses of martingales in survival analysis. They may be more surprised by case (c): with covariates, the partial likelihood ratio does

not have a unique distribution under the null, and constructing an optimal E-variable requires using the much less well-known concept of *reverse information projection (RIPr)* (Li, 1999, Li and Barron, 2000, Grünwald et al., 2019).

Contents In the remainder of this introduction, we provide a short introduction to E-variables and test martingales. In Section 2 we develop E-variables for proportional hazards without covariates, based on Cox' partial likelihood. Section 3 extends this to the case with covariates. Section 4 provides some approximations that allow us to simplify the analysis of the required sample size and the determination of E-variables that minimize it, the use of priors (although the procedure uses approximations, type-I error control remains exact) and the comparison to the classical logrank test. Section 5 provides some simulations showing the feasibility of our approach in practice, if a minimum statistical power is required. Section 6 gives a formal derivation that the Cox partial likelihood without covariates forms a nonnegative martingale. All other proofs are delegated to the appendix.

1.1 E-Variables and Test Martingales, Safety and Optimality

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 nonnegative 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 haveT 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*.

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 — see Example 4. If we used p-values rather than E-variables we 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. E-variables and test martingales can also be used to define 'always-valid confidence intervals' that remain valid under optional stopping, as outlined in Example 3.

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 Section 4.1: the GROW E-variable which maximizes, among all E-variables, $\mathbf{E}_P[\log S\langle i\rangle \mid Y\langle 0\rangle, \ldots, Y\langle i-1\rangle]$, is also the E-variable which 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 5 we provide some simulations to relate power to GROW. Note that we cannot directly use power 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 $Y\langle 0\rangle, \ldots, Y\langle i-1\rangle$.

2 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 $\vec{g} = (g_1, \ldots, g_n)$ 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 analysis, except in the formal treatment of Section 6.

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, $Y^g(t) = \sum_{j:\vec{g}_j = g} Y_j(t)$. We define $\vec{Y}(t) = (Y_1(t), \ldots, Y_n(t))$ 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.

2.1 The Simplified Process in discrete time, without censoring

In any particular realization of the setting above, we will have a sequence of event times $t\langle 1\rangle < t\langle 2\rangle < t\langle 3\rangle < \dots$ 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 th event time and $\vec{Y}\langle i\rangle = (Y_1\langle i\rangle, \dots, Y_n\langle i\rangle)$ 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 $j \in [n]$ expect 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, $\vec{Y}\langle i\rangle$ 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* $\vec{Y}\langle 0\rangle, \vec{Y}\langle 1\rangle, \dots$ with each $\vec{Y}\langle i\rangle$ 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 $\vec{g}$ and each $\theta > 0$ we define a distribution P_θ underlying this process such that:

1. $\vec{Y}\langle 0\rangle = (1, 1, \dots, 1)$, 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

$$\begin{aligned} &\text{for } j^\circ \text{ with } g_{j^\circ} = 1\text{: } P_\theta(J\langle i\rangle = j^\circ \mid Y_{j^\circ}\langle i-1\rangle = 1) := \frac{\theta}{Y^0\langle i-1\rangle + Y^1\langle i-1\rangle\theta} \\ &\text{for } j^\circ \text{ with } g_{j^\circ} = 0\text{: } P_\theta(J\langle i\rangle = j^\circ \mid Y_{j^\circ}\langle i-1\rangle = 1) = \frac{1}{Y^0\langle i-1\rangle + Y^1\langle i-1\rangle \cdot \theta}. \end{aligned} \tag{3}$$

These requirements uniquely specify P_θ. In the next subsection we shall motivate the definition (3) 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,

$$q_\theta(1 \mid (y^0, y^1)) = \frac{y^1\theta}{y^0 + y^1\theta} \text{ and } q_\theta(0 \mid (y^0, y^1)) = \frac{y^0}{y^0 + y^1\theta} \tag{4}$$

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

$$M_{\theta_1,\theta_0}\langle 0\rangle = 1 \;\; ; \;\; M_{\theta_1,\theta_0}\langle i\rangle = \frac{q_{\theta_1}(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)}{q_{\theta_0}(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)}. \tag{5}$$

By writing out the expectation, we see that

$$\mathbf{E}_{P_{\theta_0}}\left[M_{\theta_1,\theta_0}\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)\right] = \sum_{g\in\{0,1\}} q_{\theta_1}(g \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle) = 1. \tag{6}$$

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 $\vec{Y}\langle 0\rangle, \ldots, \vec{Y}\langle i-1\rangle$, and

$$M^{\langle i\rangle}_{\theta_1,\theta_0} := \prod_{j=1}^{i} M_{\theta_1,\theta_0}\langle i\rangle \tag{7}$$

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

$$\tilde{P}_{\theta_0}\left(\text{there exists } i \text{ with } M^{\langle i\rangle}_{\theta_1,\theta_0} \geq \alpha^{-1}\right) \leq \alpha. \tag{8}$$

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 will show that this is justified in Section 2.2 and 6.

Some of the examples require a generalization of M_{θ_1,θ_0} in which q_{θ_1} in (5) 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 (5) to

$$M_{r,\theta_0}\langle 0\rangle = 1 \;\;;\;\; M_{r,\theta_0}\langle i\rangle = \frac{r_i(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)}{q_{\theta_0}(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)}. \tag{9}$$

For any choice of the r_i, (6) 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 (8) 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 (9). 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 — this is further elaborated in Section 4.1. 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

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

verified to be a conditional E-variable by noting that $\mathbf{E}_{P_{\theta_0}}\left[M_{\theta_1,\theta_0}\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)\right] = 1$; see GHK for details.

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 (4) 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 illustrate this in Section 5.

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:

$$q_{W_{i+1}}(1 \mid Y^1\langle i\rangle, Y^0\langle i\rangle) = \int_\theta q_\theta(1 \mid Y^1\langle i\rangle, Y^0\langle i\rangle) w_{i+1}(\theta) d\theta = \frac{\int_\theta q_\theta(1 \mid Y^1\langle i\rangle, Y^0\langle i\rangle) \cdot \prod_{k=1}^{i} q_\theta(N^1\langle k\rangle \mid Y^1\langle k-1\rangle, Y^0\langle k-1\rangle) w(\theta) d\theta}{\int_\theta \prod_{k=1}^{i} q_\theta(N^1\langle k\rangle \mid Y^1\langle k-1\rangle, Y^0\langle k-1\rangle) w(\theta) d\theta} \tag{10}$$

and, by telescoping:

$$M^{\langle i\rangle}_{W_1,\theta_0} = \frac{\int_\theta \prod_{k=1}^{i} q_\theta(N^1\langle k\rangle \mid Y^1\langle k-1\rangle, Y^0\langle k-1\rangle) w(\theta) d\theta}{\prod_{k=1}^{i} q_\theta(N^1\langle k\rangle \mid Y^1\langle k-1\rangle, Y^0\langle k-1\rangle)} \tag{11}$$

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 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.

Example 3 [Always-Valid Confidence Sequences] Standard tests give rise to confidence intervals by varying the null and 'inverting' the corresponding tests. In analogous fashion, test martingales can be used to derive *always-valid (AV) confidence sequences*

(Darling and Robbins, 1967, Lai, 1976, Howard et al., 2018a,b). In our setting, a $(1-\alpha)$-AV confidence sequence is a sequence of confidence intervals $\{\text{CI}_i\}_{i\in\mathbb{N}}$, one for each consecutive event, such that

$$P_\theta\left(\text{there is an } i \in \mathbb{N} \text{ with } \theta \notin \text{CI}_i\right) \leq \alpha. \tag{12}$$

(see Example 4 for why we write $i \in \mathbb{N}$ rather than $i \in \{1, \ldots, Y^0\langle 0\rangle + Y^1\langle 1\rangle\}$). A standard way to design $(1-\alpha)$- AV confidence sequences is to start with a prior W on θ_1 and report, after observing i events, $\text{CI}_{i,\alpha} = [\theta_{i,L}, \theta_{i,U}]$ where $\theta_{i,L}$ is the largest θ_0 such that for all $\theta' \leq \theta_0$, $M_{W,\theta'}\langle i\rangle \geq 1/\alpha$; similarly $\theta_{i,U}$ is the smallest θ_0 such that for all $\theta' \geq \theta_0$, $M_{W,\theta'}\langle i\rangle \geq 1/\alpha$. That is, we check (11) where we vary θ_0 and we report the smallest interval such that $M_{W,\theta_0} > 1/\alpha$ outside this interval.

This will give an AV confidence sequence for arbitrary priors W — in fact it will still do so if we make the prior W a function of θ', but for simplicity we will not go into that option here.

We stress again that the AV confidence sequences give *frequentist* confidence intervals that are valid irrespective of our choice of prior W — and indeed they are different from Bayesian posterior credible intervals. The effect of the prior is that if we are 'lucky' and the data match the prior well, then the confidence intervals we end up with will be narrower than if the data contradict the prior.

Example 4 [Several Trials, Ad Lib Optional Continuation] What if we want to combine several trials, conducted in different hospitals or in different countries? In such a case we often compare a 'global' null — $\mathcal{H}_0$ is true in all trials — to an alternative that allows for different hazard ratios in different trials, with different populations. We may thus associate the k-th trial with E-variable $M_{\theta_{1,k},\theta_0}$, with $\theta_{1,k}$ varying from trial to trial — or, as in Example 2, we might use priors on θ_1 in each trial.

Suppose then that there are several trials numbered $k = 1, 2, \ldots$ and we observe subsequent events numbered $m = 1, 2, \ldots$ where $k(m)$ denotes the trial the m-th event is part of, and $i(k, m)$ is the number of events seen so far (i.e. after seeing m events in total) within trial k. Let

$$M_k\langle i\rangle := \frac{q_{\theta_k}(N_k^1\langle i\rangle \mid Y_k^1\langle i-1\rangle, Y_k^0\langle i-1\rangle)}{q_{\theta_0}(N_k^1\langle i\rangle \mid Y_k^1\langle i-1\rangle, Y_k^0\langle i-1\rangle)}$$

denote the E-variable corresponding to the i-th event in the k-th trial, with $Y_k^g\langle i\rangle$ denoting the number of people at risk in group g in trial k after i events. The evidence against H_0 after having observed s events, can then be summarized as

$$M^s_{\text{META}} := \prod_{m=1}^{s} M_{k(m)}\langle i(k(m), m)\rangle.$$

As GHK explain, in such cases the always-valid type-I error guarantee still holds: under the null, the probability that there ever comes a sequence of s events such that $M^s_{\text{META}} \geq 1/\alpha$, is still bounded by α. Thus, we effectively perform an 'on-line meta-analysis' here that remains valid irrespective of the order in which the events of the different trials come in.

Importantly, unlike in α-spending approaches, the maximum number of trials and the maximum sample size (number of events) per trial do not have to be fixed in advance; one may always decide to start a new trial, or to postpone ending a trial and wait for new data.

Example 5 [Pseudo-Staggered Entries] Consider the common case with just a single trial, in which some of the participants enter the study at a later date — as long as

whenever they enter, there is always at least one participant that enters into both groups, we can treat this scenario as a very special case of the previous one: it is as if one would start a new trial, with the same θ_1, but a new cohort of participants; one can simply combine all results pertaining to this 'new' trial with those of the previous results by multiplying the respective E-variables, again preserving type-I error guarantees. In this approach we essentially stratify the risk set by early and late entry.

A different scenario occurs when a *single* participant may enter at a later time. An asymptotic analysis of staggered entries in a sequential setting is given by Sellke and Siegmund (1983); extending it to our nonasymptotic treatment is a goal for future work — see Section 7.

2.2 Linking to Proportional Hazards without Covariates

We now consider a richer setting in which event time can be explicitly represented, yet we still discretize it to a very small (but unknown) ϵ, with $\epsilon \ll 1$ (so every time unit contains many discretized time points, each representing an interval). We will show within this setting, that if the data really come from a random process satisfying proportional hazards, i.e. for some $\theta > 0$, for all $t \geq 0$, $\lambda_{(1)}(t)/\lambda_{(0)}(t) = \theta$ where $\lambda_{(g)}$ is the hazard rate for group g, then in the limit for small ϵ, the conditional distributions (3) are correct, and thus $M^{\langle i \rangle}_{\theta_1,\theta_0}$ will truly behave like a test martingale. In Section 3 we extend this argument to Cox' proportional hazards model with covariates.

Thus, we now denote by $P_{[\epsilon],\cdot}$ a distribution for the ϵ-increment discrete-time random process $\{\vec{Y}(k \cdot \epsilon)\}_{k=0,1,2,\ldots}$. All random elements occurring inside the argument of $P_{[\epsilon],\theta}$ are defined relative to this ϵ-increment process. Thus, while we write $P_{[\epsilon],\theta}(N^g\langle i \rangle = 1 \mid \vec{Y}\langle i-1 \rangle = \vec{y})$ below it would be more correct to write $P_{[\epsilon],\theta}(N^g_{[\epsilon]}\langle i \rangle = 1 \mid \vec{Y}_{[\epsilon]}\langle i-1 \rangle = \vec{y})$ since the definition of random variables $N^g_{[\epsilon]}\langle i \rangle$ and $\vec{Y}_{[\epsilon]}\langle i-1 \rangle$ depends on ϵ; but we will generally omit this dependency in the notation.

We will generally assume that the hazard functions are Lipschitz continuous and bounded away from 0 and ∞ in t, i.e. there exists some (unknown) constants $0 < c < C < \infty$ such that for all $t \geq 0, g \in \{0,1\}$, $c < \lambda_{(g)}(t) < C$.

We say that a distribution $P_{[\epsilon]}$ for random process $\{\vec{Y}(k \cdot \epsilon)\}_{k=0,1,2,\ldots}$ as above is compatible with hazard functions $\{\lambda_j : j = 1, \ldots, n\}$ (assumed Lipschitz continuous and bounded as above) if for all $j \in [n]$, we have $P_{[\epsilon]}(Y_j(0) = 1) = 1$ and, under $P_{[\epsilon]}$ the event times $T_1, T_2, \ldots, T_n$ are i.i.d. random variables with support $\mathbb{N}$ and with marginal distribution satisfying, for $j \in [n]$,

$$\min_{(k-1)\epsilon \leq t < k\epsilon} \lambda_j(t)\epsilon \leq P_{[\epsilon]}(T_j < k\epsilon \mid (k-1)\epsilon \leq T_j) \leq \max_{(k-1)\epsilon \leq t < k\epsilon} \lambda_j(t)\epsilon. \tag{13}$$

We say that $P_{[\epsilon]}$ is *compatible* with *proportional hazards ratio θ and group assignment* $\vec{g}$, if it is compatible with hazard functions $\{\lambda_j : j = 1, \ldots, n\}$ and there exist functions $\lambda_{(1)}, \lambda_{(0)}$, again assumed Lipschitz continuous and bounded away from 0 and infinity, with, for all t, $\lambda_{(0)}(t)/\lambda_{(1)}(t) = \theta$ such that, for all j in the control group ($g_j = 0$) we have $\lambda_j = \lambda_{(0)}$ and for all j in the treatment group ($g_j = 1$) we have $\lambda_j = \lambda_{(1)}$.

We call a tuple $(i, \vec{y}, y^0, y^1)$ *compatible with i event times* if it can arise in the process of unfolding events, and the total number of event times is at least i; that is, if $i \in [n]$, y^0 is the number of 0s in $\vec{y}$, y^1 is the number of 1s in $\vec{y} \in \{0,1\}^n$, and $y^0 + y^1 = n$ if $i = 1$ and $0 < y^0 + y^1 \leq n - (i-1)$ if $i > 1$. By only requiring an inequality rather than an equality in the latter equation, we thus do allow for more than one event to happen at any time $t\langle i \rangle$ at which an event happens, but the result below implies that the probability that at the i-th event time there will be more than 1 event goes to 0 with ϵ.

Theorem 1 *Fix $\theta > 0$, and let, for all $\epsilon > 0$, $P_{[\epsilon],\theta}$ be a distribution compatible with proportional hazards ratio θ. Let $t\langle i\rangle, N^g\langle i\rangle, \vec{Y}\langle i\rangle$ be defined as in the beginning of Section 2.1. For $g \in \{0,1\}$, for all $(i, \vec{y}, y^0, y^1)$ compatible with i event times, we have:*

$$\begin{aligned} q_\theta(g \mid (y^0, y^1)) &= \lim_{\epsilon \downarrow 0} P_{[\epsilon],\theta}(N^g\langle i\rangle = 1 \mid \vec{Y}\langle i-1\rangle = \vec{y}) \\ &= \lim_{\epsilon \downarrow 0} P_{[\epsilon],\theta}(N^g\langle i\rangle = 1 \mid Y^0\langle i-1\rangle = y^0, Y^1\langle i-1\rangle = y^1) \end{aligned}$$

where $q_\theta(\cdot \mid (y^0, y^1))$ is as in (4), y^1 is the number of 1s in $\vec{y}$, and y^0 is the number of 0s. Moreover, the limits holds uniformly, e.g. for the second statement we really have $\lim_{\epsilon\downarrow 0} \sup |q_\theta(g \mid (y^0, y^1)) - P_{[\epsilon],\theta}(N^g\langle i\rangle = 1 \mid Y^0\langle i-1\rangle = y^0, Y^1\langle i-1\rangle = y^1)| = 0$ with the supremum over $g \in \{0,1\}$ and all tuples compatible with i event times.

Theorem 1 is a reformulation of existing results that underlie the standard interpretation of q_θ as a partial likelihood. The particular form we show here implies that, in the limit as $\epsilon \downarrow 0$, for all $\theta_0, \theta_1 > 0$, $\mathbf{E}_{P_{[\epsilon],\theta_0}}[M_{\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle i-1\rangle = \vec{y}] \leq 1$, so that, as long as we consider sufficiently small time scales, M_{θ_1,θ_0} will behave like an E-variable and $M^{\langle i\rangle}_{\theta_1,\theta_0}$ will behave like a test martingale, allowing for optional stopping and continuation. In Section 6 we show, using more difficult arguments, that it really *is* a nonnegative martingale if we consider continuous time directly.

2.3 Linking to Proportional Hazards without covariates, II: when ties are possible

In practice, we often observe the data at regular time intervals, which we may identify with unit time. Thus, at time $t = 1, 2, \ldots$ we observe that, between time $t-1$ and t, there have been $N^g[t-1, t]$ events in group g where $N^g[t-1, t]$ can be 0, 1, but also more than 1. In the latter case, we usually do not observe the order in which the specific events took place. We cannot represent this common situation with our previous limiting process P_θ, which requires a fully observable ordering of events. Luckily, we can define a version of a conditional distribution P_θ which is still well-defined in this situation, which is still a limit of $P_{[\epsilon],\theta}$ as defined above, and which again leads to a likelihood ratio that forms a test martingale under the null.

We consider the discrete-time random process as above, with small time steps ϵ, where we now assume that $\epsilon = 1/m$, where m is some large integer. We choose ϵ of this form to ensure that our unit time is an integer multiple of an ϵ-time interval. We first derive the relevant probability mass function, which is the analogue of $q_\theta(g|(y^0, y^1))$: fix an integer s (this will represent the number of events that happen in a given unit time interval) and let, for $v = 0, 1, \ldots, s$,

$$r_\theta(v \mid (y^0, y^1), s) := \frac{\binom{y^1}{v} \cdot \binom{y^0}{s-v} \cdot \theta^v}{\sum_{u=v_{\min}}^{v_{\max}} \binom{y^1}{u} \cdot \binom{y^0}{s-u} \cdot \theta^u}. \text{ with } v_{\min} = \max\{0, s-y^0\}; v_{\max} = \min\{s, y^1\}, \tag{14}$$

be the probability mass function of a Fisher noncentral hypergeometric distribution with parameters $(N[t-1,t], Y^1(t-1), Y^0(t-1) + Y_1(t-1), \theta) = (s, y^1, y^0 + y^1, \theta)$.

We call a tuple $(s, \vec{y}, y^0, y^1)$ *internally compatible* if it can arise in the process of unfolding events; that is, if y^0 is the number of 0s in $\vec{y}$, y^1 is the number of 1s in $\vec{y} \in \{0,1\}^n$, and if $s \leq y^0 + y^1 = n$.

Theorem 2 *Let $P_{[\epsilon],\theta}$ be a distribution compatible with proportional hazards as defined above. Let $N[K-1, K] := N^1[K-1, K] + N^0[K-1, K]$. We have for all $m \in \mathbb{N}$, all $K \in \mathbb{N}$, all internally compatible tuples $(s, \vec{y}, y^0, y^1)$ with $s \geq 1$, all $v \in \mathbb{N}_0$ with $0 \leq v \leq s$,*

that:

$$\lim_{m\to\infty} P_{[1/m],\theta}(N^1[K-1,K] = v \mid N[K-1,K] = s, \vec{Y}(K-1) = \vec{y}) = r_\theta(v \mid (y^0, y^1), s). \tag{15}$$

The same statement holds if we replace the condition $\vec{Y}(K-1) = \vec{y}$ *by* $Y^0(K-1) = y^0; Y^1(K-1) = y^1$*).*

Thus, if the data come from a process satisfying proportional hazards with rate θ, then to all intents and purposes we may act as if the number of events in group 1 in a given time unit, and given a certain number of events happening in the same time unit is hypergeometric.

As our new analogue of the E-variable M_{θ_1,θ_0}, we now define, for given $\theta_0, \theta_1 > 0$ and integer times $k = 1, 2, \ldots$:

$$U^{(K)}_{\theta_1,\theta_0} := \prod_{k=1}^{K} U(k) \ , \ U_{\theta_1,\theta_0}(k) := \begin{cases} 1 & \text{if } \ N[k-1,k] = 0 \\ \frac{r_{\theta_1}(N^1[k-1,k] \mid Y^0(k-1), Y^1(k-1), N[k-1,k])}{r_{\theta_0}(N^1[k-1,k] \mid Y^0(k-1), Y^1(k-1), N[k-1,k])} & \text{if } \ N[k-1,k] > 0. \end{cases} \tag{16}$$

Theorem 2 implies that

$$\lim_{m\to\infty} \mathbf{E}_{P_{[1/m],\theta_0}}[U_{\theta_1,\theta_0}(k) \mid \vec{Y}(k-1)] = 1.$$

Thus, in the limit for almost continuous time, $U_{\theta_1\theta_0}(k)$ becomes an E-variable conditional on $\vec{Y}(0), \vec{Y}(1), \ldots, \vec{Y}(k-1)$, and $U^{(k)}_{\theta_1,\theta_0}$ becomes a test martingale. We may thus think of $U^{(k)}_{\theta_1,\theta_0}$ being 'essentially' a test martingale; if the true process satisfies proportional hazards with rate θ_0, the type-I error of testing based on $U^{(k)}_{\theta_1,\theta_0}$ is preserved under optional stopping.

We note that, unlike M_{θ_1,θ_0}, U_{θ_1,θ_0} cannot, or at least not easily, be thought of as a likelihood under any fully defined distribution for the process $\{\vec{Y}(k\epsilon)\}_{k=0,1,2\ldots}$; it is simply set to 1 at times at which no event happens, whereas any ratio of 'real' underlying likelihoods would presumably not be 1 at all those times. But this is not a problem: for the test martingale interpretation, we merely need E-variables, a more general concept than likelihoods.

Compatibility between Theorem 1 and Theorem 2 Reassuringly, if at all times $k = 1, 2, \ldots, K$ we observed at most one event to happen inbetween time $k-1$ and k, then the evidence as measured by M and the evidence as measured by U can be reconciled. To see this, let $N(k)$ be the number of events that have happened up until time k. Then if at all times $k = 1, 2, \ldots, K$ we observed at most one event to happen inbetween time $k-1$ and k, we have $U^{(k)} = M^{\langle N(k)\rangle}$: the two processes only start to disagree once there has been a unit of time in which more than one event happpened. To see that $U^{(k)} = M^{\langle N(k)\rangle}$, note that, plugging in $v = 1, s = 1$ into (14), for $y^0, y^1 > 1$, we get:

$$r_\theta(1 \mid (y^0, y^1), 1) = \frac{\binom{y^1}{1} \cdot \binom{y^0}{0} \cdot \theta^1}{\sum_{v=0}^{1} \binom{y^1}{v} \cdot \binom{y^0}{1-v} \cdot \theta^v} = \frac{y^1 \cdot \theta}{y^0 + y^1 \cdot \theta},$$

which is the same as for the q_θ we used in the definition of the process $\{M^{\langle i\rangle}\}$; we can do an analogous derivation with $v = 0, s = 1$. Thus, both processes coincide as long as we never observe more than 1 event per time unit.

Odds vs Hazard Ratios The standard interpretation of the parameter θ in a hypergeometric distribution is as an *odds ratio*, i.e. a quantity of the form $(p_1/p_0) \cdot (1 - p_0)/(1 - p_1)$. At first sight it might seem strange that in this theorem, it takes the form $\theta = \lambda_{(1)}/\lambda_{(0)}$. What happened to the $1 - \lambda_{(g)}$'s? This is readily explained by checking the proof: at discretization level ϵ, $P_{[\epsilon],\theta}$ is really approximated by a hypergeometric with parameter $(\lambda_{(1)}\epsilon/\lambda_{(0)}\epsilon) \cdot (1 - \lambda_{(0)}\epsilon)/(1 - \lambda_{(1)}\epsilon)$. Taking the limit $\epsilon \downarrow 0$ the $1 - \lambda_{(g)}$ factors disappear.

3 Covariates: the Cox Proportional Hazard E-Variable

Fix a set of d covariates and let $\mathbf{Z} = (\vec{z}_1 \dots \vec{z}_n)$ be the matrix consisting of the covariate vectors for each participant: $\vec{z}_j = (z_{j,1}, \dots, z_{j,d})$. Just like the group assigment $\vec{g}$, we assume $\mathbf{Z}$ to be fixed (hence the covariates do not change in time), all probability measures being conditioned on its values, so that we can omit it from our notation. As in the case without covariates, we first consider the simplified process without continuous time of Section 2.1. The distribution underlying this process is now denoted $P_{\beta,\theta}$ with $\theta > 0$ and $\beta \in \mathbb{R}^d$. As before, this process is defined to satisfy the first two requirements of Section 2.1, but the third requirement now becomes: $P_{\beta,\theta}$ is given by, for $j \in [n]$ and $\vec{y} \in \{0,1\}^n$ with $\vec{y}_j = 1$:

$$P_{\beta,\theta}(J\langle i \rangle = j \mid \vec{Y}\langle i-1 \rangle = \vec{y}) := q_{\beta,\theta}(j \mid \vec{y}) := \frac{\exp(\beta^T \vec{z}_j + \theta' g_j)}{\sum_{j':\vec{y}_{j'}=1} \exp(\beta^T \vec{z}_{j'} + \theta' g_{j'})},$$

with $\theta' = \log \theta$, consistent with Cox' (1972) proportional hazards regression model: the probability that the j-th participant has an event, assuming he/she is still at risk, is proportional to the exponentiated weighted covariates, with group membership being one of the covariates. In case $\beta = 0$, this is easily seen to coincide with the definition of P_θ via (4).

We can link the new process to a more realistic process with almost-continuous time just as in Section 2.2: we let $P_{[\epsilon],\beta,\theta}$ be a distribution for the random process as in that section, and we say that $P_{[\epsilon]}$ is *compatible* with *the d-dimensional Cox proportional hazards model with parameters $\beta \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$, group assignment $\vec{g}$ and covariates $\mathbf{Z}$*, if it is compatible with hazard functions $\{\lambda_j : j = 1, \dots, n\}$ as in (13) and these functions satisfy for a function λ_{BASE} that is Lipschitz continuous and bounded away from 0 and infinity: for all t,

$$\frac{\lambda_j(t)}{\lambda_{\text{BASE}}(t)} = \exp(\beta^T \vec{z}_j + \theta g_j).$$

In any realization of the process, events will happen at times $t_1, t_2, \dots$. Let $J\langle i \rangle$ denote the set of participants with $T_j = t_i$, i.e. they suffer an event at the i-th event time. If we consider sufficiently small time scales, then the probability that more than one event happens at any even time goes to 0, and we get:

Theorem 3 *Let $P_{[\epsilon],\beta,\theta}$ a distribution compatible with Cox' proportional hazards model as above. For all $\vec{g} \in \{0,1\}^n$, all $\mathbf{Z} \in \mathbb{R}^{n \cdot d}$, all $i \in \mathbb{N}$ and $\vec{y} \in \{0,1\}^n$ with $y_j = 1$, $|\vec{y}| \leq n - (i-1)$, we have:*

$$q_{\beta,\theta}(j \mid \vec{y}) = \lim_{\epsilon \downarrow 0} P_{[\epsilon],\theta}(J\langle i \rangle = \{j\} \mid \vec{Y}\langle i-1 \rangle = \vec{y}).$$

The proof of this result is entirely analogous to the proof of Theorem 1 and is omitted from the text.

3.1 E-Variables and Martingales

Let W be a prior distribution on $\beta \in \mathbb{R}^d$ for some $d > 0$. (W may be degenerate, i.e. put mass one in a specific parameter vector β_1). We let

$$q_{W,\theta}(j \mid \vec{y}) = \int q_{\beta,\theta}(j \mid \vec{y}) dW(\beta).$$

Consider a measure ρ on $\mathbb{R}^k$ (e.g. Lebesgue or some counting measure) and we let $\mathcal{W}$ be the set of all distributions on $\mathbb{R}^k$ which have a density relative to ρ, and $\mathcal{W}^\circ \subset \mathcal{W}$ be any convex subset of $\mathcal{W}$ (we may take $\mathcal{W}^\circ = \mathcal{W}$, for example). We define $\tilde{q}_{\leftarrow W,\theta_0}(\cdot \mid \vec{y})$ to be the *reverse information projection* (Li, 1999) (RIPr) of $q_{W,\theta}(j \mid \vec{y})$ on $\{q_{W,\theta_0} : W \in \mathcal{W}^\circ\}$ such that

$$D(q_{W,\theta_1}(\cdot \mid \vec{y}) \| \tilde{q}_{\leftarrow W,\theta_0}(\cdot \mid \vec{y})) = \inf_{W' \in \mathcal{W}^\circ} D(q_{W,\theta_1}(\cdot \mid \vec{y}) \| q_{W',\theta_0}(\cdot \mid \vec{y})).$$

We know from Li (1999), Grünwald et al. (2019) that $\tilde{q}_{\leftarrow W,0}(\cdot \mid \vec{y})$ exists. As explained in the context of E-variables for 2×2 contingency tables, the fact that the random variables $Y\langle i \rangle$ constituting our random process have finite range implies that, for each W, the infimum is in fact achieved by some distribution W' with finite support on $\mathbb{R}^d$.

For given $\theta_0, \theta_1 > 0$, let

$$M_{W,\theta_1,\theta_0}\langle i \rangle = \frac{q_{W,\theta_1}(N^1\langle i \rangle \mid \vec{Y}\langle i-1 \rangle)}{q_{\leftarrow W,\theta_0}(N^1\langle i \rangle \mid \vec{Y}\langle i-1 \rangle)} \tag{17}$$

be our analogue of $M_{\theta_1,\theta_0}\langle i \rangle$ as in (5).

Theorem 4 [Corollary of Theorem 1 from GHK19] *For every prior W on $\mathbb{R}^k$, for all $\tilde{\beta} \in \mathbb{R}^d$, $\mathbf{E}_{\tilde{\beta},\theta_0}[M_{W,\theta_1,\theta_0}\langle i \rangle \mid \vec{Y}\langle i-1 \rangle)]$ is an E-variable.*

Note that the result does not require the prior W to be well-specified in any-way: under any $(\tilde{\beta}, \theta_0)$ in the null distribution, even if $\tilde{\beta}$ is completely disconnected to W, M_{W,θ_1,θ_0} is an E-variable.

How to find the RIPr While in general, it is not clear how to calculate the RIPr $q_{\leftarrow W,\theta_0}$, Li (1999), Li and Barron (2000) have designed an efficient algorithm for approximating it, which is feasable as long as we restrict $\mathcal{W}^\circ$ to be the set of all priors W for which, for all $j \in [n]$, $Q_{W,\theta_0}(J\langle i \rangle = j \mid \vec{Y}_j\langle i-1 \rangle = 1) \geq \delta$, for $\ell = 1, \ldots, k$, for some $\delta > 0$. The algorithm achieves an approximation error of $O((\log(1/\delta))/M)$ if run for M steps, where each step takes time linear in d. Since the factor is logarithmic in $1/\delta$, we can take a very small value of δ and then the requirement does not seem overly restrictive. Exploring whether the Li-Barron algorithm really allows us to compute the RIPr for the Cox model, and hence M_{W,θ_1,θ_0} in practice, is a major goal for future work.

When ordering of events is lost While in the case without covariates, our E-variables allowing for ties (several events at a time with unknown ordering) correspond to a likelihood ratio of noncentral hypergeometrics, the situation is not so simple if there are covariates — although deriving the appropriate extension of the noncentral hypergeometric partial likelihood is possible, one ends up with a hard-to-calculate formula (Peto, 1972). Various approximations have been proposed in the literature (Cox, 1972, Peto, 1972, Efron, 1974). In case these preserve the E-variable and martingale properties, they would retain type-I error probabilities under optional stopping and we could use them without problems. We do not know whether this is the case however; for the time being, we recommend handling ties by putting the events in a worst-case order, leading to the smallest values of the E-variable of interest, as this is bound to preserve the type-I error guarantees.

4 Important Approximations: Stopping Time, Bayes Predictive, Normal Likelihood

Below we present three approximate calculations that are relevant for practice. In all three cases we restrict these to the simple setting of Section 2, one-event-at-a-time and no covariates. All three require the same informally stated *Basic Condition: for both* $g \in \{0,1\}$*,* $Y^g\langle 0\rangle$ *is 'large' compared to the maximum number of events we will ever see.* Under this condition, the ratio between the number of people at risk in both groups remains approximately constant throughout the trial, so that, as a first approximation, we may approximate, for all i until the end of the trial, $q_\theta(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)$ by $q_\theta(N^1\langle i\rangle \mid Y^1\langle 0\rangle, Y^0\langle 0\rangle)$. Thus, in our calculations we will replace $P_\theta(N^g\langle i\rangle = 1 \mid Y^0\langle i-1\rangle = y^0, Y^1\langle i-1\rangle = y^1)$, defined in terms of q as in (4), by

$$P'_\theta(N^1\langle i\rangle = 1 \mid Y^0\langle i-1\rangle = y^0, Y^1\langle i-1\rangle = y^1) := q_\theta(N^1\langle i\rangle \mid Y^1\langle 0\rangle, Y^0\langle 0\rangle), \tag{18}$$

which replaces the counts at time $i-1$ by the counts at time 0. Thus, we treat the data as if it were i.i.d. Bernoulli , greatly facilitating the analysis. This Basic Condition could, for example, hold because the number of events needed to get a significant result or to stop because of futility is much smaller than the number of people at risk; or it could hold because the study will be stopped long before the number of events gets large — to give a practical example, in the ongoing Covid-19 vaccine trials, the number of people included in each trial is in the 10000s, whereas the number of events one expects to happen is in the low 100s.

4.1 Expected Stopping Time and the GROW Criterion

Let P_{θ_0} represent our null model, and let, as before, the alternative model be given as $\mathcal{H}_1 = \{P_{\theta_1} : \theta \in \Theta\}$ with $\Theta = \{\theta_1 : 0 < \theta_1 \leq \underline{\theta}_1\}$ for some $\underline{\theta}_1 < 1$. Suppose we perform a level α test based on a test martingale M_{θ,θ_0} using the aggressive stopping rule: stop as soon as $M_{\theta,\theta_0} \geq 1/\alpha$. The GROW criterion (Section 1.1, Example 1) tells us to use $\theta = \underline{\theta}_1$. Here we motivate this GROW criterion by showing that it minimizes, in a worst-case sense, the expected number of events needed before there is sufficient evidence to stop. The calculation below ignores the practical need to prepare for a bounded maximum number of events and the relation to classical statistical power. For such more complicated considerations, we need to resort to simulations as in the next section.

Recall our Basic Condition: allowing us to act as if the distribution of $Y^1\langle 1\rangle, Y^2\langle 2\rangle, \ldots$ are given by (18). This makes the random variables $N^1\langle i\rangle$ i.i.d. Bernoulli, enabling a standard argument based on Wald's (1952) identity. As said, we stop as soon as $M := M_{\theta,\theta_0} \geq 1/\alpha$ or when we run out of data, leading to a stopping time τ_θ. Suppose first that we happen to know that the data comes from a specific $\theta_1 \in \Theta_1$. Wald's identity now gives:

$$\mathbf{E}_{P'_{\theta_1}}[\tau_\theta] = \frac{\mathbf{E}_{P'_{\theta_1}}[\log M_{\theta,\theta_0}\langle\tau_\theta\rangle]}{\mathbf{E}_{P'_{\theta_1}}[\log M_{\theta,\theta_0}\langle 1\rangle]}.$$

For simplicity we will further assume that the number of people at risk is large enough compared to θ_1 so that the probability that we run out of data before we can reject is negligible. The right-hand side can then be further rewritten as

$$\frac{\mathbf{E}_{P'_{\theta_1}}[\log M_{\theta,\theta_0}\langle\tau_\theta\rangle]}{\mathbf{E}_{P'_{\theta_1}}\left[\log \frac{p'_\theta(N^1\langle 1\rangle)}{p'_{\theta_0}(N^1\langle 1\rangle)}\right]} = \frac{\log \frac{1}{\alpha} + \text{VERY SMALL}}{\mathbf{E}_{P'_{\theta_1}}\left[\log \frac{p'_\theta(N^1\langle 1\rangle)}{p'_{\theta_0}(N^1\langle 1\rangle)}\right]} \tag{19}$$

with VERY SMALL between 0 and $\log|\theta/\theta_0|$, and $p'_\theta(N^1\langle 1\rangle) = q_\theta(N^1\langle 1\rangle \mid Y^1\langle 0\rangle, Y^0\langle 0\rangle)$. The first equality is just definition, the second follows because we reject *as soon as*

$M_{\theta,\theta_0} \geq 1/\alpha$, so $M_{\theta,\theta_0}\langle\tau_\theta\rangle$ can't be smaller than $1/\alpha$, and it can't be larger by more than a factor equal to the maximum likelihood ratio at a single outcome (if we would not ignore the probability of stopping because we run out of data, there would be an additional small term in the numerator).

If we try to find the θ which minimizes this, and — as is customary in sequential analysis — we approximate the minimum by ignoring the VERY SMALL part, we see that the expression is minimized by maximizing $\mathbf{E}_{P'_{\theta_1}}[\log \frac{p_\theta(Y\langle 1\rangle)}{p_{\theta_0}(Y\langle 1\rangle)}]$ over θ. The maximum is clearly achieved by $\theta = \theta_1$; the expression in the denominator then becomes the KL divergence between two Bernoulli distributions. It follows that under θ_1, the expected number of outcomes until rejection is minimized if we set $\theta = \theta_1$. Thus, we use the GROW E-variable relative to $\{\theta_1\}$ as our actual E-variable. We still need to consider the case that, since the real $\mathcal{H}_1$ is 'composite', as statisticians, we do not know the actual θ_1; we only know $0 < \theta_1 \leq \underline{\theta}$. So we might want to take a worst-case approach and use the θ achieving

$$\max_\theta \min_{\theta_1 : 0 < \theta_1 \leq \underline{\theta}_1} \mathbf{E}_{P'_{\theta_1}}\left[\log \frac{p'_\theta(N^1\langle 1\rangle)}{p'_{\theta_0}(N^1\langle 1\rangle)}\right],$$

since, repeating the reasoning leading to (19), this θ should be close to achieving

$$\min_\theta \max_{\theta_1 : 0 < \theta_1 \leq \underline{\theta}_1} \mathbf{E}_{P'_{\theta_1}}[\tau_\theta]$$

But this just tells us to use the GROW E-variable relative to $\mathcal{H}_1$, which is what we were arguing for.

4.2 Using Bayes predictive distributions for the alternative

Calculating $M_{W,\theta_0}\langle i\rangle$ as in Example 2 and 3 amounts to calculating a Bayes predictive distribution involving an integral which is hard to evaluate for large i (at least, we did not find a prior W for which calculation is easy). One may of course approximate these predictive distributions by Gibbs sampling, but another rough-and-ready option is as follows: as in (9), let $\{r_i\}_{i\in\mathbb{N}}$ denote an arbitrary sequence of probability mass functions on $\{0,1\}$ conditioned on integers (y^0, y^1). By the reasoning underneath (9), $M_{r,\theta_0}\langle i\rangle$ also gives a conditional E-variable and its product over time gives a test martingale — we simply replaced q_{θ_1} by a sequence of distributions that possibly lie outside our model.

Suppose we have a minimum clinically relevant θ_1 in mind and we ideally would like to use q_W with W a prior peaked at θ_1. Directly calculating M_{W,θ_0} is not straightforward, but we may approximate q_W in the numerator by the following r_i: let $y_0^g = Y^0\langle g\rangle$. We first determine $p_1 = y_0^1\theta_1/(y_0^0 + y_0^1\theta_1)$. This is the probability of the first event falling in group 1 according to θ_1. We now take V_m to be a beta-distribution with density v_m on Bernoulli parameter $\mu \in [0,1]$ that peaks at p_1: $v_m(\mu) \propto \mu^{mp_1}(1-\mu)^{m(1-p_1)}$; the larger the value of m, the sharper the peak around p_1. We now set r_i to be the Bayes predictive distribution after observing $N^1\langle 1\rangle, \ldots, N^1\langle i-1\rangle$, based on prior V_m. Analogously to (10):

$$r_{i+1}(1 \mid Y^1\langle i\rangle, Y^0\langle i\rangle) = \int_\mu \mu dW(\mu \mid N^1\langle 1\rangle, \ldots, N^1\langle i\rangle) = \frac{\sum_{k=1}^i N^1\langle k\rangle + mp_1 + 1}{i + m + 2}. \quad (20)$$

We can easily turn this into an always-valid confidence sequence, replacing $M_{W,\theta'}$ in the definition below (12) by $M_{r,\theta'}$ based on r_i as in (20). We stress that the use of r_i instead of q_W does not compromise on safety: type-I errors and confidence sequences based on $M_{\rho,\theta'}$ remain valid.

4.3 Comparison to Normal Likelihood for the Logrank Statistic

The traditional, nonsequential treatment of the logrank test considers data of a fixed sample size and analyzes the logrank statistic at that sample size. Schoenfeld (1981)

shows that, under the null, this statistic is asymptotically normally distributed (which was essentially already stated by Cox (1972)), and can be used to perform fixed-sample size power analyses. This raises the question whether our E-variable $M_{\theta_1,\theta_0}\langle i\rangle$ can also be related to a likelihood ratio between normal densities defined on a 'local' logrank statistic (not for the full sample, but just for the i-th event). A priori it is not at all clear whether Schoenfeld's asymptotic, fixed sample result has a nonasymptotic sequential counterpart, but it turns out that for hazard ratios close to 1 (also noted by Cox (1972)) and at sample sizes with total number of events much smaller than the number of participants at risk in either group, there is a strong correspondence after all.

Thus, we are looking to check if

$$M_{\theta_1,\theta_0}\langle i\rangle = \frac{q_{\theta_1}(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)}{q_{\theta_0}(N^1\langle i\rangle \mid Y^1\langle i-1\rangle, Y^0\langle i-1\rangle)} \approx \frac{q_{\theta_1}(N^1\langle i\rangle \mid Y^1\langle 0\rangle, Y^0\langle 0\rangle)}{q_{\theta_0}(N^1\langle i\rangle \mid Y^1\langle 0\rangle, Y^0\langle 0\rangle)} \text{ is close to } M'_{\mu_1,\mu_0}\langle i\rangle := \frac{\phi_{\mu_1,1}(Z\langle i\rangle)}{\phi_{\mu_0,1}(Z\langle i\rangle)} \quad (21)$$

for some μ_1 and μ_0 to be determined below as functions of θ_1 and θ_0, where $\phi_{\mu,\sigma}$ is the density of a normally distributed RV with mean μ and variance σ^2. Here the $\approx$ equality is already justified by our Basic Condition, and we will now check wither the latter two quantities are close as well under this condition. $Z(i)$ is defined as the standard logrank statistic that would be observed based *only* on the i-th event:

$$Z\langle i\rangle = \frac{N^1\langle i\rangle - P^1\langle i\rangle}{\sqrt{P^1\langle i\rangle \cdot P^0\langle i\rangle}} \text{ with } N^1\langle i\rangle \in \{0,1\},$$

and $P^g\langle i\rangle = Y^g\langle i-1\rangle/(Y^0\langle i-1\rangle + Y^1\langle i-1\rangle)$ the fraction of people at risk in group g when the i-th event happens.

Now, Schoenfeld (1981) shows that asymptotically, the distribution of the logrank statistic based on all events so far,

$$\bar{Z}\langle 1:i\rangle := \frac{\sum_{k=1}^{i}(N^1\langle k\rangle - P^1\langle k\rangle)}{\sqrt{\sum_{k=1}^{i} P^1\langle k\rangle \cdot P^0\langle k\rangle}} \quad (22)$$

converges to a normal with variance 1 and mean given by $\sqrt{P^1\langle 0\rangle P^0\langle 0\rangle i'}\log\theta$ where i' is approximately equal to the number of events i. Now under our Basic Condition that the number of people at risk in both groups is a lot larger than the number of events that we will ever measure, we have $\bar{Z}\langle 1:i\rangle \approx (1/\sqrt{i})\cdot\sum_{k=1}^{i} Z\langle k\rangle$. This suggests (but does not prove) that $Z\langle i\rangle$ itself can be approximated by a normal distribution with mean $\sqrt{P^1\langle 0\rangle P^0\langle 0\rangle}\log\theta$. We will thus take the μ_g in (21) to be $\sqrt{P^1\langle 0\rangle P^0\langle 0\rangle}\log\theta_g$.

Now, simulations show that with this choice, M_{μ_1,μ_0} in (21) gets extremely close to M_{θ_1,θ_0} as long as θ_1 is close to 1 and the number of people at risk in both groups is much larger than number of events. In Figure 1 we plot $M_{\theta_1,1}\langle i\rangle/M'_{\mu_1,0}\langle i\rangle$ as a function of $\log\theta_1$ (the plots are identical for the case that the i-th event is in group 1 and the case that it is in group 0, and are identical for different i, as long as the number of people at risk in both groups is the same). For example, suppose we test a hazard ratio of $\theta_1 = 0.8$, corresponding to a Schoenfeld mean of $\mu_1 = -0.11$ against $\theta_0 = 1$ ($\mu_0 = 0$). At this level the approximation is extremely tight: the ratio of the Bernoulli and normal likelihood ratios is 1.000013. If we consider hazard ratio $\theta_1 = 0.5$, we get a ratio of 1.001. This is small, but may not be completely negligible any more — if we have a 1000 events, the joint likelihood ratio can be expected to be off by a factor of about $(1+.001)^{1000} \approx \exp(1)$ — at least, because also the ratio between the number of persons at risks would change in a real experiment. Note that as the hazard ratio moves further away from 1, the error made by the normal approximation *per event* can become quite large after all — on the

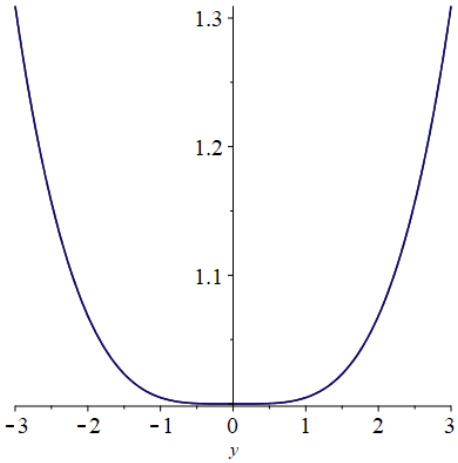

Figure 1: Schoenfeld Normal vs Bernoulli

other hand, less events are needed to reject the null. In the next section we shall see that therefore, in practice, the Gaussian approximation of the likelihood works quite well, even for moderately extreme θ as long as we aim for power that is not too high (say 0.8).

5 Some Simulations

In this section we investigate by simulation the results and the discussion from the previous sections.

Recall from our discussion in Section 2.1 that the process M_{θ_1,θ_0} does not depend on the event times themselves, but only on their ranks: in which order the events happened in each group. Furthermore, Theorem 1 allows us to simulate efficiently the order in which the events of the survival process happen. Indeed, 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 $\alpha = 0.05$. As in Section 4.1, 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 $1/\alpha$. We interpret the infimum of an empty set to be ∞, so that $\tau_{\theta_1} = \infty$ if the threshold is never crossed. 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 estimated experimentally 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 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 2 we show the expected value of the random number of events τ_{θ_1}, and of its stopped version $\tau_{\theta_1} \wedge i_{\max}$ under the null hypothesis.

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-testi, 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 as described underneath (22); this is a standard classical approach. 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, is of comparable size. For small hazard ratios Schoenfeld's Gaussian approximation deteriorates, and determining the sample size needed for achieving a power of 80% when using a z-test as above is overly optimistic. To account for this, for small hazard ratios, in Figure 2, we computed by simulation the exact sample size needed to achieve power 80% when using the z-test based on the log rank statistic, and for larger hazard ratios we used the sample size based on the correctness of the Gaussian approximation.

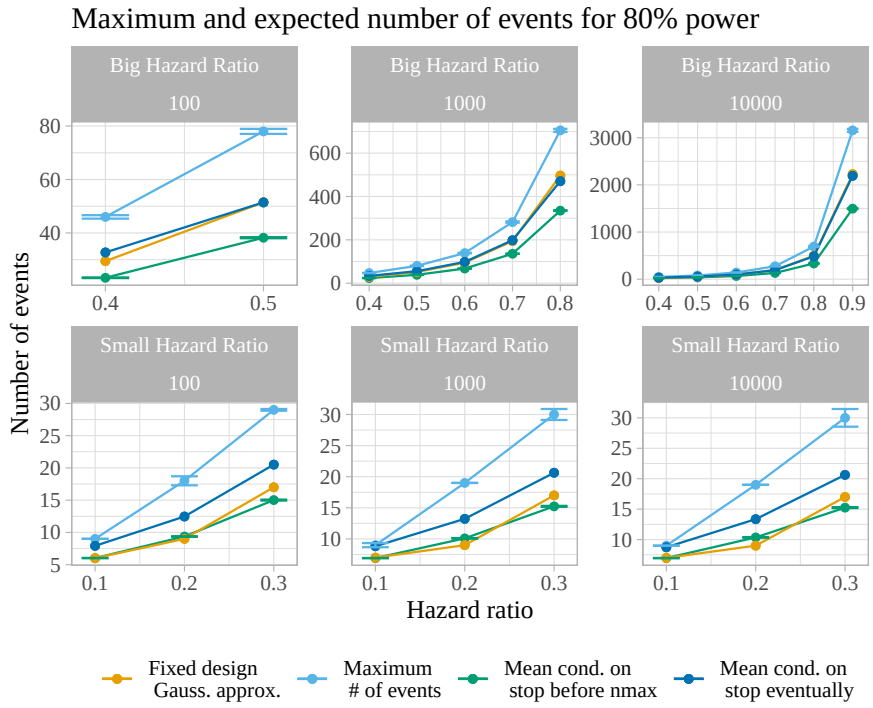

Figure 2: Each panel is a different starting number of subjects in group 1, and the same number of patients in group 0. In general, one is expected to stop earlier than the value $i_{\max}$

5.1 Gaussian approximation

In Figure 3 we show the relation between the maximum number of events $i_{\max}$ derived from our test martingale $M_{\theta_1,1}$, and its "Gaussian approximation" $M'^{\langle i \rangle}_{\theta_1,1}$ defined in (21). Note that the Gaussian approximation of the Bernoulli likelihood is not expected to be accurate for small values of θ_1. We note that in the case when we start with 10000 patients in both groups, it is necessary to have a maximum number of events ~ 1000 for $\theta_1 = 0.9$ to retain 80% power. This is not regime where we conjecture that the "Gaussian" approximation to the Bernoulli likelihood process is valid. We performed experiments with larger starting group sizes, where we witnessed the "Gaussian" approximation stopping earlier than the exact test, as explained in Section 4.3.

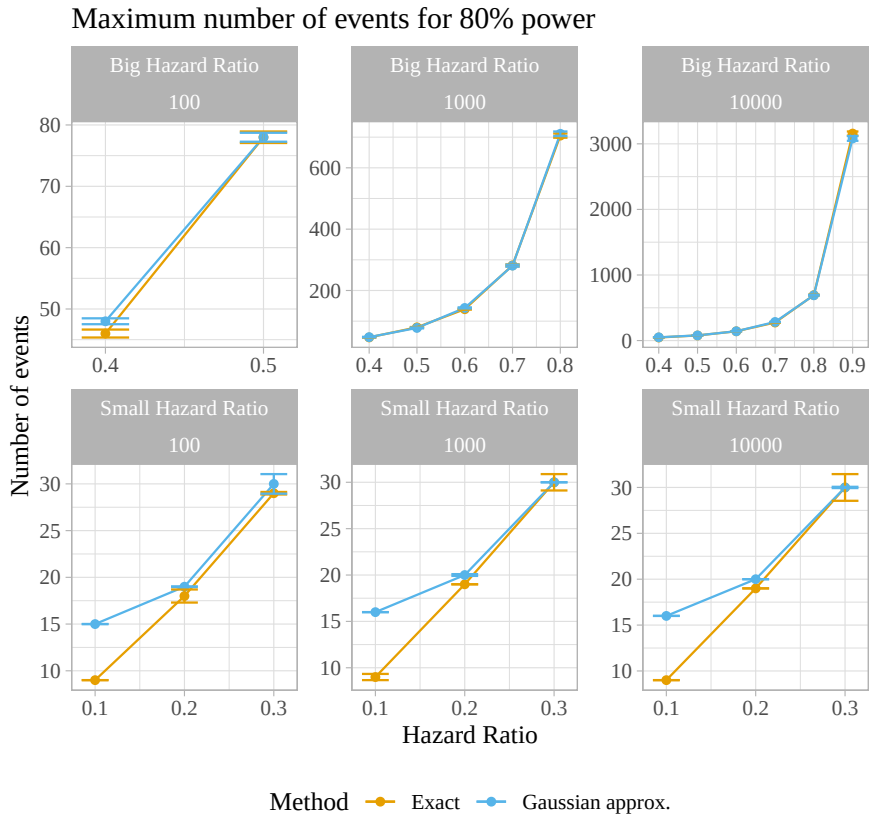

Figure 3: Stopping times resulting from using the exact martingale $M^{\langle i \rangle}_{\theta_1,1}$ and its 'Gaussian' approximation $M'^{\langle i \rangle}_{\theta_1,1}$ defined in (21) (but not with fixed design). Each panel shows the result of simulating using a different starting number of patients (100, 1000, and 10000), which is equal in both groups. For small starting number of patients, it is not possible to have error 80% and stop before we run out of patients in the experiments.

5.2 Misspecification, and beta prior on Θ_1

As we noted earlier, it may happen that data come from a distribution with a more extreme hazard ratio than we anticipated. As we argued in Section 4.2, the best choice (the one that leads to the smallest stopping time τ_{θ_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 Figure 4 we illustrate such a situation when we start with 1000 in both groups. We generated data using different hazard ratios, and used a 'misspecified' $M^{\text{miss}}_{\theta_1,1}$ that always used $\theta_1 = 0.8$. Note that while this is still the GROW (minimax optimal) martingale test under $\mathcal{H}_1 = \{P_\theta : \theta_1 \leq 0.8\}$, if we knew the true θ_1, we could obtain a faster-growing test martingale. We estimated the maximum number of events $i^{\text{miss}}_{\max}$ that allows for 80% power. We compare this to three other alternatives: first, the maximum number of events $i_{\max}$ that we would have obtained had we known real value of θ, and the $i^{\text{Beta}}_{\max}$ that we obtain by using the Beta prior as described in Section 4.2 with $m = 5$, and $m = 100$. Recall that for larger values of m, the beta prior described in 4.2 concentrates at θ_1.

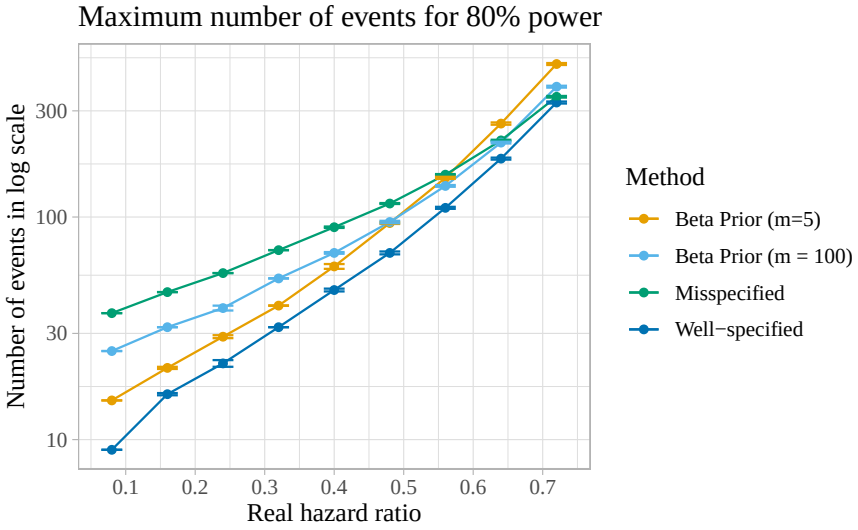

Figure 4: We show the number of events at which one can stop retaining 80% power using the process $M_{\theta_1,0}$ with $\theta_1 = 0.80$ when the real hazard ratio was different. We used a starting number of patients equal to 1000 in both groups. Using a Beta prior with parameter $m = 5$ can lead to earlier stopping when the real one hazard ratio is close to the real compared to using a mispecified test. Larger values of m give a bahavior increasingly similar to using the misspecified $M_{\theta_1,0}$. Placing a prior on Θ_1 can lead to earlier stopping than using a misspecified test in some situations.

6 Formal Proof that M_{θ_0,θ_1} is a test martingale in continuous time

In this section we give a formal proof that the process M_{θ_0,θ_1} presented in Section 2 is a martingale in continuous time, and under noninformative right censoring. First we need a battery of definitions, and preliminary results. Our main reference in this section is the text of Fleming and Harrington (2011).

Let $\theta = \lambda_{(1)}(t)/\lambda_{(0)}(t)$ be the hazard ratio, which is assumed to be a constant in time under the proportional hazards ratio model. We observe data for n participants. For the j-th participant, we observe the triplet (Z_j, X_j, δ_j), where each element is as follows. Z_j, Y_j, Y^g are defined as before. We also observe $X_j = \min\{C_j, T_j\}$, the minimum between the random censoring time C_j, and the time T_j when the event of interest occurred. Thirdly, we observe the indicator δ_j of whether the event of interest occured

($\delta_j = 1$ in that case), or not ($\delta_j = 0$). Thus, the data for participants consists of n independent triplets $(Z_1, X_1, \delta_1), \ldots, (Z_n, X_n, \delta_n)$.

We assume that the event times $T_1, \ldots, T_n$ are absolutely continuous random variables that are iid conditionally on which group each participant belongs to, and that the censoring times $C_1, \ldots, C_n$ are iid with an arbitrary distribution independent of that of the event times.

We now let $N_j(t) = \mathbf{1}_{X_j \leq t, \delta_j = 1}$, and $N_j^C(t) = \mathbf{1}_{X_j \leq t, \delta_j = 0}$ the processes that count whether or not the event of interest or censoring already happened at time t for participant j. Let $\mathcal{F}_t$ be the filtration generated by these two processes $\mathcal{F}_t = \sigma\{N_j(s), N_j^C(s) : 0 \leq s \leq t\}$. Under these assumptions, Fleming and Harrington (2011, Theorem 1.3.1) show that for each $j = 1, \ldots, n$, the process $t \mapsto N_j(t) - \lambda_j(t)$ is a martingale adapted to the filtration $\mathcal{F}_t$, where $\lambda_j(t) = Y_j(t)\lambda_{(1)}(t)$ if $Z_j = 1$, and $\lambda_j(t) = Y_j(t)\lambda_{(0)}(t)$ if $Z_j = 0$.

We continue with a few more definitions. For $j = 1, \ldots, n$, let $q_{\theta,j}(t)$ be the likelihood of randomly sampling a participant from the group Z_j at odds $\theta Y^1(t) : Y^0(t)$ at time t, that is,

$$q_{\theta,j}(t) = Z_j \frac{\theta Y^1(t)}{Y^0(t) + \theta Y^1(t)} + (1 - Z_j)\frac{Y^0(t)}{Y^0(t) + \theta Y^1(t)}.$$

Let θ_0 and θ_1 be pair of hazard ratios under the null and under the alternative hypothesis, respectively. Define $r_j(t)$ as the probability ratio between the probabilities $q_{\theta_1,j}(t)$ and $q_{\theta_0,j}(t)$ defined in the previous display, that is,

$$r_j(t) = \frac{q_{\theta_1,j}(t)}{q_{\theta_0,j}(t)} = Z_j \frac{\theta_1}{\theta_0} \frac{Y^0(t) + \theta_0 Y^1(t)}{Y^0(t) + \theta_1 Y^1(t)} + (1 - Z_j)\frac{Y^0(t) + \theta_0 Y^1(t)}{Y^0(t) + \theta_1 Y^1(t)}.$$

Let $(T_1, j_1), \ldots, (T_K, j_K)$ be the pairs of times T_k at which the event of interest in either group are observed, and their corresponding patient index j_k. Define the process $M(t)$ by $M(0) = 1$ and by

$$M(t) = \prod_{T_k \leq t} r_{j_k}(T_k)$$

for $t > 0$, where we interpret the empty product as being equal to 1. We can rewrite this process in terms of stochastic integrals with respect counting processes $N_1, \ldots, N_n$

$$M(t) = \exp\left(\sum_{j=1}^{n} \int_0^t \ln(r_j(t))\mathrm{d}N_j(t)\right), \tag{23}$$

where as before $N_i(t) = \mathbf{1}_{X_i \leq t}$, and the integration is performed pathwise.

With these definitions and preliminary results, we are in position to formulate the main result of this section.

Theorem 5 *The process $t \mapsto M(t)$ defined in (23) is a martingale adapted to the filtration $\{\mathcal{F}_t\}_{t>0}$ given by $\mathcal{F}_t = \sigma\{N_j(s), N_j^C(s) : i = 1, \ldots, n, 0 \leq s \leq t\}$.*

Proof: Fleming and Harrington (2011, Theorem 1.5.1) shows that it is sufficient to write $M(t)$ as $M(t) = \sum_{j=1}^n \int_0^s H_j(s)(\mathrm{d}N_j(s) - \mathrm{d}\lambda_j(s))$ for predictable processes $H_1(s), \ldots, H_n(s)$. We will show in the following that in fact M can be written as

$$M(t) = \sum_{j=1}^{n} \int_0^t M(s_-)(r_j(s) - 1)(\mathrm{d}N_j(s) - \lambda_j(s)\mathrm{d}t). \tag{24}$$

The result will follow because for $j = 1, \ldots, n$, the processes $t \mapsto M(s_-)(r_j(s) - 1)$ are bounded, and since they are left continuous, they are predictable (see Fleming and Harrington, 2011, Lemma 1.4.1).

From (23), we can see that $M(t)$ satisfies*

$$M(t) = \sum_{i=1}^{n} \int_0^t M(s_-)(r_{\Theta,i}(s) - 1)\mathrm{d}N_i(s)$$

thus, comparing the previous display and our goal (24), it is sufficient to prove that

$$\sum_i (r_i(t) - 1)\lambda_i(t) = 0. \tag{25}$$

To this end, computation shows that

$$r_i(t) - 1 = Z_i \left(\frac{\theta_1 - \theta_0}{\theta_0} \frac{Y^0(t)}{Y^0(t) + \theta_1 Y^1(t)} \right) + (1 - Z_i) \left((\theta_0 - \theta_1) \frac{Y^1(t)}{Y^0(t) + \theta_1 Y^1(t)} \right). \tag{26}$$

Multiply by λ_i and sum up to obtain that

$$\sum_i (r_i(t) - 1)\lambda_i(t) = \left(\frac{\theta_1 - \theta_0}{\theta_0} \frac{Y^0(t)}{Y^0(t) + \theta_1 Y^1(t)} \right) \sum_{i:Z_i=1} \lambda_i(t) + \left((\theta_0 - \theta_1) \frac{Y^1(t)}{Y^0(t) + \theta_1 Y^1(t)} \right) \sum_{i:Z_i=0} \lambda_i(t). \tag{27}$$

Recall that $\lambda_i(t) = Y_i(t)(Z_i\lambda_{(1)}(t) + (1 - Z_i)\lambda_{(0)}(t))$ so that $\sum_{i:Z_i=1} \lambda_i(t) = Y^1(t)\lambda_{(1)}(t)$ and $\sum_{i:Z_i=0} \lambda_i(t) = Y^0(t)\lambda_{(0)}(t)$. Using these last observations and the fact that under the null hypothesis $\lambda_{(1)}(t)/\lambda_{(0)}(t) = \theta_0$, the previous display implies that (25) holds. This implies that $M(t)$ satisfies (24) and that consequently it is a martingale, as claimed. □

7 Conclusion and Future Work

We introduced the safe logrank test, a version of the logrank test that can retain type-I error guarantees under optional stopping and continuation. We gave an extension to Cox' proportional hazards regression which seems very promising since it provides type-I error guarantees even if the alternative model is equipped with arbitrary priors. In future work, we plan to implement this extension — which requires the use of sophisticated methods for estimating mixture models.

Earlier approaches to sequential time-to-event analysis were also studied under scenarios of staggered entry, where each patient has its own event time (e.g. time to death since surgery), but patients do not enter the follow-up simultaneously (such that the risk set of e.g. a two-day-after-surgery event changes when new participants enter and survive two days). Sellke and Siegmund (1983) and Slud (1984) show that, in general, martingale properties cannot be preserved under such staggered entry, but that asymptotic results are hopeful (Sellke and Siegmund, 1983) as long as certain scenarios are excluded (Slud, 1984). When all participants' risk is on the same (calendar) time scale (e.g. infection risk in a pandemic), or new patients enter in large groups (allowing us to stratify — Example 5), staggered entry poses no problem for our methods. But research is still ongoing into those scenarios in which our inference is fully safe for patient time under staggered entry, and those that need extra care.

*If $X(t) = \int_0^t H(s)\mathrm{d}N(s)$ for a predictable process H and a counting process N, then for $x \mapsto F(x)$ a continuous function it holds that $\mathrm{d}F(X(t)) = (F(X(t_-) + H(t)) - F(X(t_-)))\mathrm{d}N(t)$

8 Acknowledgements

This work is part of the research programme with project number 617.001. 651, which is financed by the Dutch Research Council (NWO).

References

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.

Bradley Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1974. with discussion.

Thomas R. Fleming and David P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

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.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Exponential line-crossing inequalities. *arXiv:1808.03204 [math]*, August 2018a. URL `http://arxiv.org/abs/1808.03204`. arXiv: 1808.03204.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, non-parametric, non-asymptotic confidence sequences. *arXiv:1810.08240 [math, stat]*, October 2018b. URL `http://arxiv.org/abs/1810.08240`. arXiv: 1810.08240.

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

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.

J.Q. Li and A.R. Barron. Mixture density estimation. In S.A. Solla, T.K. Leen, and K-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, Cambridge, MA, 2000. MIT Press.

Qiang (Jonathan) Li. *Estimation of Mixture Models*. PhD Thesis, Yale University, New Haven, CT, USA, 1999.

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.

Luigi Pace and Alessandra Salvan. Likelihood, replicability and Robbins' confidence sequences. *International Statistical Review*, 2019.

Richard Peto. Discussion on the paper 'Regression models and Life Tables by Sir David R. Cox. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34 (2):205–208, 1972.

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.

Eric V. Slud. Sequential Linear Rank Tests for Two-Sample Censored Survival Data. *Annals of Statistics*, 12(2):551–571, June 1984. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346505. URL `https://projecteuclid.org/euclid.aos/1176346505`. Publisher: Institute of Mathematical Statistics.

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

Abraham Wald. *Sequential Analysis*. Wiley, New York, 1952.

A Proof of Theorem 1

Let $(i, \vec{y}, y^0, y^1)$ be compatible wit i event times. Then there exists a $k \in \mathbb{N}$ (in fact, there exist infinitely many k) such that the following conditional probability is well-defined, and we can derive:

$$\begin{aligned}
&P_{[\epsilon],\theta}(N^1\langle i\rangle = 1 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon) = \\
&P_{[\epsilon],\theta}(N^1\langle i\rangle = 1, N^0\langle i\rangle = 0 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon) + \\
&\quad P_{[\epsilon],\theta}(N^1\langle i\rangle = 1, N^0\langle i\rangle > 0 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon) = \\
&(1+f(\epsilon))\cdot P_{[\epsilon],\theta}(N^1\langle i\rangle = 1, N^0\langle i\rangle = 0 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon) \qquad (28)
\end{aligned}$$

where $f(\epsilon)$ is given by

$$\frac{P_{[\epsilon],\theta}(N^1\langle i\rangle = 1, N^0\langle i\rangle > 0 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon)}{P_{[\epsilon],\theta}(N^1\langle i\rangle = 1, N^0\langle i\rangle = 0 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon)} =$$
$$\frac{\binom{y^1}{1}\cdot\lambda_{(1)}(t)\rho_{\epsilon,(1),t}^{y^1-1}\cdot\left(\sum_{1\le j\le y^0}\binom{y_0}{j}\cdot(\lambda_{(0)}(t))^j\epsilon^j\rho_{\epsilon,(0),t}^{y^0-j}\right)}{\binom{y^1}{1}\cdot\lambda_{(1)}(t)\rho_{\epsilon,(1),t}^{y^1-1}\rho_{\epsilon,(0),t}^{y^0}}(1+o(\epsilon)) = o'(\epsilon) \tag{29}$$

where $t = k\epsilon$, we abbreviated $1-\epsilon\lambda_{(g)}(t)$ to $\rho_{\epsilon,(g),t}$, and o and o' are both 'small o' functions that go to 0 faster than ϵ (the limit implicit in $o'(\epsilon)$ holds because of the factor ϵ^j in the numerator). Also, with the same abbrevations,

$$P_{[\epsilon],\theta}(N^1\langle i\rangle = 1, N^0\langle i\rangle = 0 \mid \vec{Y}\langle i-1\rangle = \vec{y}, t\langle i\rangle = k\cdot\epsilon) =$$
$$\frac{\binom{y^1}{1}\cdot\lambda_{(1)}(t)\rho_{\epsilon,(1),t}^{y^1-1}\rho_{\epsilon,(0),t}^{y^0}}{\binom{y^1}{1}\cdot\lambda_{(1)}(t)\rho_{\epsilon,(1),t}^{y^0-1}\rho_{\epsilon,(0),t}^{y^0}+\binom{y^0}{1}\cdot\lambda_{(0)}(t)\rho_{\epsilon,(0),t}^{y^0-1}\rho_{\epsilon,(1),t}^{y^1}}\cdot(1+o''(\epsilon))$$
$$= \frac{y^1\theta}{y^0+y^1\theta}\cdot(1+o'''(\epsilon)) \tag{30}$$

where again $t = k\epsilon$ and o'' and o''' are two more 'small o' functions that go to 0 faster than ϵ. We see that the limit does not depend on k, so that the first limit result of Theorem 1 follow for $g = 1$ in (28); the proof for $g = 0$ is entirely analogous. The second limit of Theorem 1 follows by repeating exactly the same reasoning with the conditioning event $\vec{Y}\langle i-1\rangle = \vec{y}$ replaced by $Y^1\langle i-1\rangle = y^1, Y^0\langle i-1\rangle = y^0$. Uniformity of the limits follows directly by noting that the supremum is over a finite set.

B Proof of Theorem 2

We define the m-component binary vector

$$\vec{N}_K := (N[K, K+\epsilon], N[K+\epsilon, K+2\epsilon], \ldots, N[K+(m-1)\epsilon, K+1])$$

which indicates at what ϵ-width time intervals between time K and $K+1$ events happened in either group; the kth entry of this vector is the number of events that happened between time $K+(k-1)\epsilon$ and $K+k\epsilon$. Similarly, $\vec{N}_K^g$ is a defined as a vector whose kth entry indicates the number of events in group g inbetween time $K+(k-1)\epsilon$ and $K+k\epsilon$.

We now first determine, for binary m-component vectors $\vec{s}$ with $s \ge v$ ones, the probability

$$P_{[\epsilon],\theta}(N^1[K-1,K] = v \mid \vec{N}_{K-1} = \vec{s}) =$$
$$\frac{\sum_{\vec{w}\in\{0,1\}^m:|\vec{w}|=v} P_{[\epsilon],\theta}(\vec{N}_{K-1}^1 = \vec{w}, \vec{N}_{K-1} = \vec{s})}{\sum_{v'=0,1,\ldots,s}\sum_{\vec{w}\in\{0,1\}^m:|\vec{w}|=v'} P_{[\epsilon],\theta}(\vec{N}_{K-1}^1 = \vec{w}, \vec{N}_{K-1} = \vec{s})}. \tag{31}$$

After having found an expression for this conditional probability, we will argue that it coincides, up to a $(1+o(\epsilon))$ factor, with the probability that we are after. We can write the probabilities occurring in the sums in (31) as

$$P_{[\epsilon],\theta}(\vec{N}_{K-1}^1 = \vec{w}, \vec{N}_{K-1} = \vec{s}) = a_{\vec{w}}\cdot a'_{\vec{w}},$$

where $a_{\vec{w}}$, satisfying (32), collects all factors corresponding to intervals of length ϵ during which nothing happens (i.e. the components of $\vec{s}$ with 0 entry), and $a'_{\vec{w}}$, given by (33), collects the s other factors:

$$\left((1-\bar{\lambda}_1\epsilon)^{y^1}(1-\bar{\lambda}_0\epsilon)^{y^0}\right)^{m-s} \le a_{\vec{w}} \le \left((1-\underline{\lambda}_1\epsilon)^{y^1-s}(1-\underline{\lambda}_0\epsilon)^{y^0-s}\right)^{m-s}. \tag{32}$$

where $\bar{\lambda}_g = \sup_{K \le t < K+1} \lambda_{(g)}(t)$, $\underline{\lambda}_g = \inf_{K \le t < K+1} \lambda_{(g)}(t)$, so that $a_{\vec{w}} = 1 + o(\epsilon)$.

To get an expression for $a'_{\vec{w}}$, first let, for $u \in [s]$, ℓ_u be the index of the u-th component in $\vec{s}$ that is 1. That is, if $\vec{s} = (0, 0, 0, 1, 0, 1, \ldots)$ then $\ell_1 = 4$, $\ell_2 = 6$, and so on. $\vec{N}_{K-1} = \vec{s}$ thus means that a single event happened in intervals $[(K-1) + 3\epsilon, (K-1) + 4\epsilon]$ and $[(K-1) + 5\epsilon, (K-1) + 6\epsilon]$, and so on. Let $k_1(\vec{v}), k_2(\vec{v}), \ldots, k_s(\vec{v})$ be defined such that $k_u(\vec{v}) = 1$ if $v_{\ell_u} = 1$ and $k_u(\vec{v}) = 0$ otherwise. $k_u(\vec{v})$ represents whether the u-th of the s events inbetween time $K-1$ and K was in group 1 or 0. Below we abbreviate $\lambda_{(g)}((K-1) + \ell_u\epsilon)$ to $\lambda_{g,u}$. We can then write:

$$a'_{\vec{w}} = \prod_{u=1}^{s} (y^1 - \sum_{u' < u} k_{u'}(\vec{w}))(1 - \lambda_{1,u}\epsilon)^{y^1 - \sum_{u' \le u} k_{u'}(\vec{w})} \cdot (\lambda_{1,u}\epsilon)^{k_u(\vec{w})} \cdot$$
$$(y^0 - \sum_{u' < u} (1 - k_{u'}(\vec{w})))(1 - \lambda_{0,u}\epsilon)^{y^0 - \sum_{u' \le u}(1 - k_{u'}(\vec{w}))} \cdot (\lambda_{0,u}\epsilon)^{1 - k_u(\vec{w})} = (1 + o(\epsilon))p(\vec{w}), \tag{33}$$

with, with $w = |\vec{w}|$, and using that, by our proportional hazards assumption, $\lambda_{1,u} = \theta\lambda_{0,u}$:

$$\begin{aligned} p(\vec{w}) &= \prod_{u=1}^{s} (y^1 - \sum_{0 < u' < u} k_{u'}(\vec{w})) \cdot \lambda_{1,u}^{k_u(\vec{w})} \cdot (y^0 - \sum_{0 < u' < u} (1 - k_{u'}(\vec{w}))) \cdot \lambda_{0,u}^{1 - k_u(\vec{w})} \\ &= \prod_{u=0}^{w-1} (y^1 - u) \cdot \theta^w \cdot \prod_{u=0}^{s-w-1} (y^0 - u) \prod_{u=1}^{s} \lambda_{0,u} = w!(s-w)! \binom{y^1}{w} \binom{y^0}{s-w} \theta^w \prod_{u=1}^{s} \lambda_{0,u}. \end{aligned}$$

This gives:

$$\sum_{\vec{w} \in \{0,1\}^m : |\vec{w}| = w} p(\vec{w}) = \binom{s}{w} p(\vec{w}) = s! \cdot \binom{y^1}{w} \binom{y^0}{s-w} \theta^w \prod_{u=1}^{s} \lambda_{0,u}$$

Plugging this into (31), we get

$$P_{[\epsilon],\theta}(N^1[K-1, K] = v \mid \vec{N}_{K-1} = \vec{s}) = (1 + o(\epsilon)) \cdot \frac{\sum_{\vec{w} \in \{0,1\}^m : |\vec{w}| = v} p(\vec{w})}{\sum_{v' = 0,1,\ldots,s} \sum_{\vec{w} \in \{0,1\}^m : |\vec{w}| = v'} p(\vec{w})}$$
$$= (1 + o(\epsilon)) \cdot \frac{s! \cdot \binom{y^1}{v}\binom{y^0}{s-v} \theta^v \prod_{u=1}^{s} \lambda_{0,u}}{\sum_{v' = 0,1,\ldots,s} s! \cdot \binom{y^1}{v'}\binom{y^0}{s-v'} \theta^{v'} \prod_{u=1}^{s} \lambda_{0,u}} = (1 + o(\epsilon)) \cdot \frac{\binom{y^1}{v}\binom{y^0}{s-v} \theta^v}{\sum_{v' = 0..s} \binom{y^1}{v'}\binom{y^0}{s-v'} \theta^w}$$

which only depends on the number of events s inbetween $K = 1$ and K and not on their specific location in the vector $\vec{s}$. We have thus shown that

$$P_{[\epsilon],\theta}(N^1[K-1, K] = v \mid N[K-1, K] = s, \mathcal{E}_K) =$$
$$P_{[\epsilon],\theta}(N^1[K-1, K] = v \mid \vec{N}_{K-1} = \vec{s}) = (1 + o(\epsilon)) \cdot \frac{\binom{y^1}{v}\binom{y^0}{s-v} \theta^v}{\sum_{v' = 0..s} \binom{y^1}{v'}\binom{y^0}{s-v'} \theta^w}$$

where $\mathcal{E}_K$ is the event (in our measure space) that not more than one event took place in any interval between $K-1$ and K. The result then follows by standard rewriting of the above probability using that $\Pr(\mathcal{E}_K) = 1 - o(\epsilon)$, as can be shown in a way similar to (29); we omit the details.