The Anytime-Valid Logrank Test: Error Control Under Continuous Monitoring with Unlimited Horizon

JUDITH TER SCHURE*, MURIEL F. PÉREZ-ORTIZ*, ALEXANDER LY, AND
PETER D. GRÜNWALD

Abstract

We introduce the anytime-valid (AV) logrank test, a version of the logrank test that provides type-I error guarantees under optional stopping and optional continuation. The test is sequential without the need to specify a maximum sample size or stopping rule, and allows for cumulative meta-analysis with type-I error control. The method can be extended to define anytime-valid confidence intervals. The logrank test is an instance of the martingale tests based on E-variables that have been recently developed. We demonstrate type-I error guarantees for the test in a semiparametric setting of proportional hazards, show explicitly how to extend it to ties and confidence sequences and indicate further extensions to the full Cox regression model. Using a Gaussian approximation on the logrank statistic, we show that the AV logrank test (which itself is always exact) has a similar rejection region to O'Brien-Fleming α-spending but with the potential to achieve 100% power by optional continuation. Although our approach to *study design* requires a larger sample size, the *expected* sample size is competitive by optional stopping.

KEYWORDS AND PHRASES: Alpha saving, Exact, Interim analysis, Optional stopping, Proportional hazards.

1. INTRODUCTION

The logrank test is arguably the most important tool for the statistical comparison of time-to-event data between two groups of participants. Our main focus is when the two groups refer to the treatment and control groups in a randomized controlled trial; the outcome of interest are event times, that is, the time elapsed until an outcome of interest. The logrank test, in turn, uses a simplified version of the proportional hazard ratio model of Cox [3]. For a fixed sample size and under this model, Cox gave a simple but profound insight: inference can be performed using the partial likelihood of having observed the events in the particular order that they were observed. To this end, the logrank test [20, 24], the score test associated to the Cox' partial likelihood, is optimal for fixed sample size and a restricted alternative. Large-sample properties of the logrank test are known in very general settings [41, 32, 1]. Nevertheless, it is clear that the fixed-sample regime can be overly restrictive. Indeed, due to ethical and practical constraints in human survival-time medical trials, interim analyses may be performed to terminate the study earlier than planned if needed. Consequently, it has been of fundamental importance to develop methods for the sequential analysis of time-to-event data in general; for the logrank test, in particular.

In order to legitimate the use of sequential boundary decisions, asymptotic approximations over the study period have been developed for the logrank statistic [42, 36, 38]. The results in this line of work show the convergence of the sequentially computed logrank statistic to a rescaled Brownian motion under very general censoring and participant-arrival patterns. When interim analyses are only performed at discrete times, the decision boundaries based on continuously monitoring the logrank statistic are known to be overly conservative. This deficiency is addressed by group-sequential and α-spending methods, which, using knowledge of the interim analysis times relative to a predefined maximum number of events, allow for tighter decision boundaries [25, 22, 14]. These sequential methods allow several interim looks at the data to stop for efficacy (if the treatment shows to be beneficial) or futility (if the study is no longer likely to reach statistical significance).

Despite the profound impact that these methods have had in statistical practice, the requirement of a maximum sample size limits the utility of a promising but nonsignificant study once the maximum sample size is reached. Because of their design, extending such a trial makes it impossible to control their type-I error. Moreover, the evidence gathered in new—possibly unplanned—trials cannot be added in a typical retrospective meta-analysis, when the number of trials or timing of the meta-analysis are dependent on the trial results. Such dependencies introduce accumulation bias and invalidate the assumptions of conventional statistical procedures in meta-analysis [34]. In order

*These authors share first authorship. Correspondence concerning this article may be addressed to Peter Grünwald, Email address: pdg@cwi.nl.

to address these deficiencies, we look for flexible anytime-valid methods that provide type-I error control in two situations: (1) optional stopping, which refers to halting the experiment earlier or later than planned under arbitrary stopping rules, and (2) meta-analysis and optional continuation, which refers to the aggregation of evidence of possibly interdependent studies. Just as the existing methods, our approach is connected to early work by H. Robbins and collaborators [5, 16]. Most notably, existing approaches come with fixed stopping rules, which are not desirable in the use cases that are of our present interest. The details of the present approach are very different, and to some extent, as we will see, more straightforward.

The main result of this work is the anytime-valid (AV) logrank test, an anytime-valid test for the statistical comparison of time-to-event data from two groups of participants. The AV logrank test uses the exact ratio of the sequentially computed Cox partial likelihood as test statistic. This is in contrast to the conventional logrank test, that can be interpreted as transforming a test statistic into a p-value whose distribution is, in all applications we are aware of, not determined exactly but rather approximated by a normal or a generalized beta distribution [24]. Such approximations are valid only in asymptotic senses, and even then, some further conditions must often hold. In contrast to these approximations, the exact anytime-valid logrank test has guaranteed type-I error control also for small sample sizes, without worries whether asymptotic approximations to the sampling distribution are justified. The advantage of having an exact test becomes particularly clear in the case of unbalanced allocation, when both control and treatment groups start with different numbers of participants. For this case, it has been documented that α-spending approaches do not provide strong type-I error guarantees due to the approximations involved [51]. The basic version of the AV logrank test is, however, exact; unbalanced allocation presents no difficulties.

From a technical point of view, we show, under general patterns of incomplete observation, that under the composite null hypothesis our test statistic is a continuous-time martingale with expected value equal to one. Statistics with this sequential property are referred to as test martingales; they form the basis of anytime-valid tests [28]. The AV logrank test is a concrete instance of such a test martingale derived from the recent theory of anytime-valid hypothesis testing based on E-processes [29, 9, 37, 46]. At each time an event takes place, it takes on the form of a likelihood ratio ratio between two multinomial distributions in a sampling-without-replacement setting. Together with the underlying assumption of continuous time, this without-replacement property makes the test martingale and corresponding filtration different in spirit than in most existing work (such as Lindon and Malek [19] who consider discrete test martingales for multinomial experiments under sampling-with-replacement). Nevertheless, just as in [48, 49]

who also considered a sampling without-replacement (but otherwise quite different) setting, we can show similar results on anytime-validity as in discrete-time, sampling-with-replacement settings.

In contrast to p-values, an analysis based on E-processes can extend existing trials as well as inform the decision to start new trials and meta-analyses, while still controlling type-I error rate. Type-I error control is retained even (i) if the E-process is monitored continuously and the trial is stopped early whenever the evidence is convincing, (ii) if the evidence of a promising trial is increased by extending the experiment and (iii) if a trial result spurs a new trial with the intention to combine them in a meta-analysis.

The AV logrank test was developed with a specific application in mind and it illustrates its usefulness. Some of the authors were involved in applying the AV logrank test to the continuous meta-analysis of seven Coronavirus disease (COVID-19) clinical trials—the results are available as a living systematic review including code and summary data to reproduce the analysis [33]. This analysis was performed concurrently with the trials in a so-called Anytime Live and Leading Interim (ALL-IN) meta-analysis [35]. We remark that even in the presence of dependencies between the existence and size of the trials, the test based on the multiplication of the values of E-processes retains type-I error control as long as all trials test the same global null hypothesis, as was the case in the above application. This is generally useful if we want to combine the results of several trials in a bottom-up retrospective meta-analysis, where 'bottom-up retrospective' means that the trials did not happen in coordination, and the decision to group them together was made after they had already started, so that no coordinated 'top-down' stopping rule can be enforced. It is even possible to obtain an interim meta-analysis result by combining *interim* results of *ongoing* trials by multiplication, stepping beyond the realm of existing sequential approaches.

1.1 Contributions and Outline

We begin with Section 2, where we review the special instance of Cox' proportional hazards model for the two-group setting. There, we set the assumptions and notation used in the rest of the article. The definitions presented there are standard. In Section 3, we define the AV logrank test and prove that is indeed anytime valid. We first do this for (a) the case with only a group indicator (no other covariates) and without simultaneous events (ties). There, we also discuss its optimality properties and extend it to (b) the case with ties and to (c) the case when one wants to learn the actual effect size from the data. As usual for anytime-valid tests, they keep providing nonasymptotic type-I error control even if the alternative is wildly misspecified. This is of particular interest for the important case in which the null expresses 'both groups are indistinguishable', highlighted in Example 3.2.

These results hinge on showing that the likelihood underlying Cox' proportional hazards model can be used to define *E*-variables and test martingales. In Section 4, we show a Gaussian approximation to the AV logrank statistic that is useful in the common situation when only summary statistics are available. We then provide extensive computer simulations to compare the AV logrank test to the classic logrank test and α-spending approaches. In Section 4.1, we show that the exact AV logrank test has a similar rejection region to O'Brien-Fleming α-spending for those designs and hazard ratios where it is well-approximated by a Gaussian AV logrank test. While always needing a small amount of extra data in the design phase (the price for indefinite optional continuation), the expected sample sizes needed for true rejections remain very competitive. During the design phase of a study, we might want to design for a maximum sample size in order to achieve a certain power, but need a smaller sample size on average during the study since we can safely engage in optional stopping. In Section 5, we show that AV-logrank-type tests can be combined through multiplication to perform meta-analysis, and in Section 6, we show how the test can be used to derive confidence sequences for the hazard ratio. In Section 7, we compare the sample sizes that are needed during the design phase in order to achieve a targeted power. Throughout the paper, we present several different instantiations of the AV logrank test. In the final Section 8 we provide recommendations for which one to use in which practical situation. There we also make concluding remarks and discuss future research directions, in particular the possibility to include covariates other than group membership, i.e. the full Cox model, which is theoretically straightforward but practically challenging unless the number of covariates remains small.

We remark that once the definitions are in place, the technical results are mostly straightforward consequences from earlier work; in particular, of the work of Cox [4], Slud [39] and Andersen et al. [1]. The novelty of the present work is thus mainly in *defining* the AV logrank test and showing by computer simulation that, while being substantially more flexible, it is competitive with existing approaches—the classic logrank test with fixed design and in combination with α-spending.

Next to the main body of this article, we provide two appendices. We delegate to Appendix A proofs and remarks that, while important, are not needed to follow the main development. Most importantly, the particular E-variable we design is *growth-rate optimal in the worst case*, GROW (see Section 3.1). Grünwald et al. [9] provide several motivations for this criterion; we provide an additional one using an argument of Breiman [2], which does not seem to be widely known. This argument shows a connection between growth-rate optimality and tests with minimal expected stopping time. In Appendix B, we discuss additional details of the extension to the case of general covariates.

2. PROPORTIONAL HAZARDS MODEL AND COX' PARTIAL LIKELIHOOD

We begin by describing the hypothesis that is being tested, the data that are available, and Cox' proportional hazards model. We are interested in comparing the survival rates between two groups of participants, Group *A* and Group *B*. In a randomized controlled trial, Group *A* would signify the control group; Group *B*, the treatment group. We assume that the available data about m participants are a subset of $\{(X^i, g^i, \delta^i) : i = 1, \ldots, m\}$, where $X^i = \min\{T^i, C^i\}$ is the minimum between the event time T^i and the (possibly infinite) censoring time C^i; g^i is a zero-one covariate depending on group membership ($g^i = 0$ signifies that $i \in A$; $g^i = 1$, that $i \in B$); and $\delta^i = \mathbf{1}\{X^i = T^i\}$ is the indicator of whether the event was witnessed before censoring or not. The idea is that at (continuous) time t, we have observed all (and only) those tuples (X^i, g^i, δ^i) with $X^i \le t$, i.e. those participants that have experienced an event. Let m^A be the number of members of Group *A* and m^B the number of members of Group *B*—then $m^A + m^B = m$. Define $\mathbf{g} = (g^1, \ldots, g^m)$, the vector of group memberships. We assume that $T^1, \ldots, T^m, C^1, \ldots, C^m$ are independent and have continuous distribution functions. The continuity assumption precludes tied observations; we relax this assumption later on, in Section 3.3. For $i = 1, \ldots, m$, the survival rates are quantified by the hazard functions $\lambda^i = (\lambda^i_t)_{t \ge 0}$ for T_i, given by

$$\lambda^i_t = -\frac{\mathrm{d}}{\mathrm{d}t} \ln \mathbf{P}\{T^i \ge t\}. \tag{2.1}$$

As is customary, the hazard function λ^i at t can be interpreted via the conditional probability of witnessing an event in a short time span provided that the event has not been witnessed up to t, that is,

$$\mathbf{P}\{t \le T^i < t + \Delta t \mid t \le T^i\} = \lambda^i_t \Delta t + o(\Delta t) \quad \text{as} \quad \Delta t \to 0. \tag{2.2}$$

Given our interest in comparing the survival rates between the two groups, suppose that all participants i of Group *A* have a common hazard function $\lambda^i_t = \lambda^A_t$; members i of Group *B*, $\lambda^i_t = \lambda^B_t$ (thus, for all participants i within the same group, the event times T_i are identically distributed). Using the data, we wish to test proportional hazards hypotheses. Concretely, we test the hypotheses $\mathcal{H}_0$ that the hazard function of the members of both groups satisfy $\lambda^A_t = \theta_0 \lambda^B_t$, against an alternative hypothesis $\mathcal{H}_1$ that $\lambda^B_t = \theta \lambda^A_t$ for a $\theta \ne \theta_0$. As a first application of the methods that we develop, we consider a left-sided alternative, that is,

$$\mathcal{H}_0 : \lambda^B_t = \theta_0 \lambda^A_t \quad \text{vs.} \quad \mathcal{H}_1 : \lambda^B_t = \theta \lambda^A_t$$
$$\text{for some} \quad \theta \le \theta_1 < \theta_0 \text{ and all } t, \tag{2.3}$$

where θ is known as the hazard ratio and is the main quantity of statistical interest, and θ_1 would be, in a clinical trial,

a minimal clinically relevant effect size. The alternative is what we hope for in case of negative events, such as death, with treatments that are set out to lower (relative to the control condition) the hazard rate. Notice that the hypotheses in (2.3) are, in fact, nonparametric. Similarly, if the event is positive, e.g., recovery from an infection, we would typically set a right-sided alternative, which can be also be treated with the present methods. Two-sided alternatives and also the full alternative hypothesis $\mathcal{H}_1' : \theta \neq \theta_0$ are also amenable to the methods that will follow. We remark, however, that all the methods retain their type-I error guarantees irrespective of the specific alternative that we use. As will be seen in Example 3.2 below, this is of particular interest in the simple, yet important case with $\theta_0 = 1$ and full alternative $\mathcal{H}_1' : \theta \neq 1$—in that case, we can think of our test as really testing 'the treatment has no effect' against general alternative 'the treatment has an effect' (even though the proportional hazard assumption may be substantially violated).

We now turn to defining Cox' partial likelihood PL_t, which is at the center of our approach. To that end, we need a battery of standard definitions—we lay them out to establish the notation. Let $y_t^i = \mathbf{1}\{X^i \geq t\}$ be the at-risk process, that is, the indicator of whether participant i is still at risk at time t, and let $\bar{y}_t^A = \sum_{i \in A} y_t^i$ and $\bar{y}_t^B = \sum_{i \in B} y_t^i$ be the number of participants at risk in each of the groups at time t. Define $\mathbf{y}_t = (y_t^1, \ldots, y_t^m)$, the vector of at-risk processes, and $\mathcal{R}_t = \{j : y_t^j = 1\}$, the set of participants at risk at time t. Let $T^{(1)} < T^{(2)} < \cdots < T^{(\bar{N}_\infty)}$ be the set of ordered events times that were witnessed (not censored). Note that, if all participants witness the event and censoring is absent, $\bar{N}_\infty = m$. For each $k = 1, \ldots, \bar{N}_\infty$, let $I_{(k)}$ be the index of the individual that witnessed the event at time $T^{(k)}$. This means, for example, that if participant with label three was the fifth to witness the event, then $I_{(5)} = 3$. Abbreviate by $y_{(k)}^i, \bar{y}_{(k)}^A, \bar{y}_{(k)}^B, \mathcal{R}_{(k)}$ the corresponding quantities at event time $T^{(k)}$, and define $g^{(k)} := g^{I_{(k)}}$. Cox' partial likelihood $\mathrm{PL}_{\theta,t}$ can be sequentially computed by

$$\mathrm{PL}_{\theta,t} = \prod_{k:T^{(k)} \leq t} \frac{\theta^{g^{(k)}}}{\sum_{l \in \mathcal{R}_{T^{(k)}}} \theta^{g^l}} = \prod_{k:T^{(k)} \leq t} \frac{\theta^{g^{(k)}}}{\bar{y}_{(k)}^A + \theta \bar{y}_{(k)}^B}. \quad (2.4)$$

Cox' likelihood evaluated at the event times $T^{(1)}, T^{(2)}, \ldots$ coincides to that of a sequence of multinomial trials where, at event time $T^{(k)}$, each of the participants $i \in \mathcal{R}_{(k)}$ witnesses the event with probability

$$p_{\theta,(k)}(\,i\,) := \mathbf{P}\{I_{(k)} = i \mid \mathbf{y}_{(l)}, \mathbf{g}; \ l = 1, \ldots k\},$$
$$p_{\theta,(k)}(\,i\,) = \frac{\theta^{g^i}}{\bar{y}_{(k)}^A + \theta \bar{y}_{(k)}^B}. \quad (2.5)$$

We note that if, under the null, the groups are indistinguishable (see also Example 3.2 below), $\theta = \theta_0 = 1$, (2.4) reduces to the likelihood of a sequence of draws-with-replacement

from an urn that initially contains m^A marbles of color A and m^B marbles of color B:

$$\mathrm{PL}_{\theta,t} = \prod_{k:T^{(k)} \leq t} \frac{\theta^{g^{(k)}}}{\bar{y}_{(k)}^A + \bar{y}_{(k)}^B}. \quad (2.6)$$

Cox showed that, indeed, conditionally on all the information accrued strictly before $T^{(k)}$, the probability that participant i observes an event at time $T^{(k)}$ is exactly $p_{\theta,(k)}(\,i\,)$ as long as the hazard ratio is θ. With these likelihood computations at hand, we are in place to show the main contribution of this article, the AV logrank test, which uses the partial likelihood ratio as the test statistic.

3. THE AV LOGRANK TEST

In this section the AV logrank test for (2.3) is introduced; its type-I error guarantees and optimality properties are investigated. We give a solution to the first of the purposes laid down in the introduction: we show that the AV logrank test is anytime valid—its type-I error guarantees are not affected by optional stopping. The fact that it is also type-I-error-safe under optional continuation, our second purpose, is proven in Section 5. Without further ado, we define the AV logrank statistic $S_{\theta_0,t}^{\theta_1}$, typically, $\theta_0 = 1$, for (2.3) as the partial likelihood ratio

$$S_{\theta_0,t}^{\theta_1} = \frac{\mathrm{PL}_{\theta_1,t}}{\mathrm{PL}_{\theta_0,t}} = \prod_{k:T^{(k)} \leq t} \frac{p_{\theta_1,(k)}(I_{(k)})}{p_{\theta_0,(k)}(I_{(k)})}. \quad (3.1)$$

Here, $p_{\theta,(k)}$ is as defined in (2.5); the product that defines our statistic $S_{\theta_0,t}^{\theta_1}$ runs over the events that have been witnessed up to and including time t, and the empty product is taken to be equal to one. As is conventional with likelihood ratios, high values of $S_{\theta_0,t}^{\theta_1}$ are indicative that the alternative hypothesis is better than the null hypothesis at the describing the data. Given a tolerable type-I error bound α and an arbitrary random time τ, the AV logrank test is the test that rejects the null hypothesis if $S_{\theta_0\tau}^{\theta_1}$ is above the threshold $1/\alpha$, that is,

$$\xi_{\theta_0,\tau}^{\theta_1} = \mathbf{1}\{S_{\theta_0,\tau}^{\theta_1} \geq 1/\alpha\} := \begin{cases} 1 & \text{if } S_{\theta_0,\tau}^{\theta_1} \geq 1/\alpha \\ 0 & \text{otherwise.} \end{cases} \quad (3.2)$$

As we will see, by its sequential properties, $S_{\theta_0,t}^{\theta_1}$ takes large values with small probability under the null hypothesis uniformly over time, which translates into type-I error control for the test $\xi_{\theta_0,\tau}^{\theta_1}$. This observation is behind the any-time validity of the AV logrank test, and of anytime-valid tests in general (more details and general constructions to the effect of anytime-valid sequential testing can be found in the work of Ramdas et al. [28]). We show in the following proposition that the test $\xi_{\theta_0,\tau}^{\theta_1}$ has the desired type-I error control.

Proposition 3.1. *Let* $\mathbf{P}_0$ *be any distribution under which the hazard ratio is equal to* θ_0, *and let* τ *be any random time. The test* $\xi_{\theta_0,\tau}^{\theta_1} = \mathbf{1}\{S_{\theta_0,\tau}^{\theta_1} \geq 1/\alpha\}$, *where* $S_{\theta_0,t}^{\theta_1}$ *is as in (3.1), has level* α, *that is,*

$$\mathbf{P}_0\{\xi_{\theta_0,\tau}^{\theta_1} = 1\} \leq \alpha.$$

This result can be readily obtained using the sequential-multinomial interpretation of Cox' likelihood ratio. As we will see, in Section 3.1, this result can be interpreted in terms of *E*-variables and *E*-processes [9]. Define the process $(S_{\theta_0,(k)}^{\theta_1})_{k=1,2,\dots}$ as the value of the AV logrank statistic at the event times $T^{(k)}$, that is, $S_{\theta_0,(k)}^{\theta_1} := S_{\theta_0,T^{(k)}}^{\theta_1}$. In this time discretization, the AV logrank statistic is the product of random variables

$$R_{\theta_0,(k)}^{\theta_1} = p_{\theta_1,(k)}(I_{(k)})/p_{\theta_0,(k)}(I_{(k)}), \qquad (3.3)$$

the one-outcome partial likelihood ratio for the *k*th event, where $p_{\theta_0,(k)}$ is as in (2.5) and $k = 1, 2, \dots$.

Proof of Proposition 3.1. Under any distribution under which the hazard ratio is θ_0, the fact that the likelihood of observing $I_{(k)}$ conditionally on $\{\mathbf{y}_{(l)} : l = 1, \dots, k\}$ equals $p_{\theta_0,(k)}(I_{(K)})$ implies that

$$\mathbf{E}[R_{\theta_0,(k)}^{\theta_1} \mid \mathbf{y}_{(1)}, \dots, \mathbf{y}_{(k)}] = \sum_{j \in \mathcal{R}_{(k)}} p_{\theta_0,(k)}(j) \frac{p_{\theta_1,(k)}(j)}{p_{\theta_0,(k)}(j)} = 1.$$
$$(3.4)$$

This immediately shows that $S_{\theta_0,(k)}^{\theta_1} = \prod_{i \leq k} R_{\theta_0,(k)}^{\theta_1}$ is a test martingale, a nonnegative martingale with expected value equal to one, with respect to the filtration $\mathcal{F}_- = (\mathcal{F}_{(k)-})_{k=1,2,\dots}$ of sigma-algebras $\mathcal{F}_{(k)-} = \sigma(\mathbf{y}_{(k)} : k = 1, \dots, k)$. Next, the type-I error control for the the test $\xi_{\theta_0}^{\theta_1}$ follows from Ville's inequality, which asserts that, under the null hypothesis, the test martingale $S_{\theta_0,(k)}^{\theta_1}$ takes large values with small probability. Ville's inequality [44] implies that

$$\mathbf{P}\left\{ \sup_{k=1,2,\dots} S_{\theta_0,(k)}^{\theta_1} \geq 1/\alpha \right\} \leq \mathbf{E}[S_{\theta_0,(1)}^{\theta_1}]\alpha = \alpha.$$

The previous display is a bound on ever making a type-I error when using the AV logrank test $\xi_{\theta_0,\tau}^{\theta_1}$. □

Proposition 3.1 is a variation of the standard statement and result establishing that tests based on E-processes are anytime-valid [29]. As remarked by a referee, one notable difference is that, within a clinical trial setting, in the standard statement each subsequent factor in $S_{\theta_0,t}^{\theta_1}$ would refer to an additional participant coming in, whereas in our setting, it corresponds to an event, meaning that a patient *leaves*. Nevertheless, the technique used to prove Proposition 3.1 is entirely analogous to the proof of the standard result (namely, Ville's inequality), as has previously shown to be also the case in other anytime-valid work that deals, like ours, with a drawing-without-replacement setting [48, 49].

Under general patterns of incomplete observation—like independent censoring or independent left truncation—, the AV logrank test provides the same type-I error guarantees. To prove this we do need to go beyond the standard, discrete-time setting: we give an alternative proof of Proposition 3.1 in Appendix A using the counting-process formalism [1]. There, we show that if the compensators of the underlying counting processes have a certain general product structure—which is the case under complete observation—, the AV logrank test is anytime-valid. We then refer to Andersen et al. [1], who show that this structure is preserved under said patterns of incomplete observation.

The AV-logrank test is optimal—in a sense to be defined in the next section—among a large family of statistics. A second look at the proof of Proposition 3.1 suggests a generalization of the AV logrank statistic given in (3.1). Let, for each k, $q_{(k)}$ be a probability distribution on participants in the risk set $\mathcal{R}_{(k)}$ which is only allowed to depend on $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(k)}$. Analogously to (3.3), we define the one-outcome ratio $R_{\theta_0,(k)}^q := q_{(k)}(I_{(k)})/p_{\theta_0,(k)}(I_{(k)})$—we now use $q_{(k)}$ instead of p_{θ_1}—, and

$$S_{\theta_0,t}^q := \prod_{k:T^{(k)} \leq t} R_{\theta_0,(k)}^q = \prod_{k:T^{(k)} \leq t} \frac{q_{(k)}(I_{(k)})}{p_{\theta_0,(k)}(I_{(k)})}. \qquad (3.5)$$

A modification of the previous argument shows, for any random time τ, a type-I error guarantee for the test $\xi_{\theta_0,\tau}^q$ based on the value of $S_{\theta_0,\tau}^q$, that is, $\xi_{\theta_0,\tau}^q := \mathbf{1}\{S_{\theta_0,\tau}^q \geq 1/\alpha\}$ (see Proposition 3.1). Any such test is also anytime valid as long as each $q_{(k)}$ depends on the data only through $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(k)}$. In Section 3.2, we use this generalization to provide tests when no value of θ_1 is available. This generalization raises a natural question about the optimality of the AV logrank test based on (3.1) among test statistics of the form (3.5). This is the subject of the next section.

3.1 E-variables and Optimality

The random variables $\{R_{\theta_0,(k)}^{\theta_1}\}_{k=1,2\dots}$ from (3.3) and $\{R_{\theta_0,(k)}^q\}_{k=1,2\dots}$ from (3.5) are examples of (conditional) *E*-variables—nonnegative random variables whose (conditional) expected value is below 1 under every distribution in the null hypothesis. *E*-variables and *E*-processes are the "correct" generalization of likelihood ratios to the case that either or both $\mathcal{H}_0$ and $\mathcal{H}_1$ are composite and can be interpreted in terms of gambling [9, 37, 28]. Under this gambling interpretation, a test martingale, a product of conditional *E*-variables, is the total profit made in a sequential gambling game where no earnings are expected under the null hypothesis. The analogy is thus between profit and evidence: no evidence can be gained against the null hypothesis if it is true. Just as *p*-values, the definition of *E*-variables and test martingales does not need any mention of an alternative hypothesis. However, if a composite set of alternative distributions is available, a gambler who is skeptical of the

null distribution might want to maximize the speed of evidence accumulation (or of capital growth) under the alternative hypothesis. The worst-case growth rate is defined (conservatively) as the smallest expectation of the logarithm of the E-variable under the alternative. Consequently, any E-variable achieving it is called GROW, for Growth-Rate Optimal in the Worst case (see the work of Grünwald [9] and Shafer [37] for additional reasons to use this optimality criterion).

We instantiate this reasoning to our present problem. For the left-sided alternative (2.3), the choice $R^{\theta_1}_{\theta_0,(k)}$ is conditionally GROW because it maximizes the worst-case conditional growth rate

$$R^q_{\theta_0,(k)} \mapsto \min_{\theta \leq \theta_1} \mathbf{E}_\theta[\ln R^q_{\theta_0,(k)} | \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}],$$

over all valid choices of $q_{(k)}$ (which can only depend on the data through $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}$), that is,

$$\min_{\theta \leq \theta_1} \mathbf{E}_\theta[\ln R^{\theta_1}_{\theta_0,(k)} | \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}]$$
$$= \max_q \min_{\theta \leq \theta_1} \mathbf{E}_\theta[\ln R^q_{\theta_0,(k)} | \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}].$$

In Appendix A.1, we show that in the limit that the risk sets are much larger than the number of events that are witnessed, this worst-case growth criterion yields a test that minimizes the worst-case expected stopping time—under the alternative hypothesis—among the tests that stop as soon as $S^q_{\theta_0,t} \geq 1/\alpha$. Thus, among all possible AV logrank tests of the form (3.5), there are strong reasons to choose $\xi^{\theta_1}_{\theta_0,\tau}$.

In a similar fashion, a test can be constructed for two sided alternatives. Indeed, consider a testing problem of the form

$$\mathcal{H}_0 : \lambda^B = \lambda^A \quad \text{vs.}$$
$$\mathcal{H}_1 : \lambda^B = \theta\lambda^A \text{ for some } \theta \leq \theta_1 \text{ or } \theta \geq 1/\theta_1, \quad (3.6)$$

where $\theta_1 < 1$. For this problem, we can create a weighted, conditionally GROW, E-variable by using $R^{2-\text{sided}} = \frac{1}{2}R^{\theta_1}_{\theta_0,(k)} + \frac{1}{2}R^{1/\theta_1}_{\theta_0,(k)}$.

3.2 Learning the Hazard Ratio from Data

So far, the alternative hypotheses that we have studied are of the form $\mathcal{H}_1 : \theta \leq \theta_1$ for some value of $\theta_1 < 1$. In some cases, such a value of θ_1 is available from the context of the analysis. For instance, θ_1 can correspond to a minimal clinically relevant effect that is satisfactory in a medical trial. However, sometimes it is not clear which value θ_1 to chose. Still, statistics of the form (3.5) are useful to test a null hypothesis $\mathcal{H}_0$ as in (2.3). Indeed, for each k, we can use conditional probability mass functions $q_{(k)}$ that depend on data observed on $t < T^{(k)}$ and enable us to implicitly learn the hazard ratio θ. Two standard types of such alternatives are readily available [29]: a prequential plug-in

likelihood and a Bayes predictive distribution. We proceed to describe the former, on which we base our experiments. For completeness we also briefly describe the latter in Appendix A.4.

Using only the data observed in $t < T^{(k)}$, let $\breve{\theta}_{(k)}$ be the smoothed maximum likelihood estimator

$$\breve{\theta}_{(k)} = \arg\max_{\theta \geq 0} \left(p_{\theta,0} \times \prod_{k':T^{(k')}<T^{(k)}} p_{\theta,(k')}(I_{(k')}) \right), \quad (3.7)$$

where $p_{\theta,0}$ is a smoothing based on the likelihood of having observed two "virtual" data points prior to the observed data, that is, $p_{\theta,0} = 1/(\bar{y}_0^A + 1 + \theta(\bar{y}_0^B + 1)) \times \theta/(\bar{y}_0^A + \theta(\bar{y}_0^B + 1))$. The statistic $S^{\text{preq}}_{\theta_0,t}$ is (3.5) with $q_{(k)} = p_{\breve{\theta}_{(k)},(k)}$, and it can also be used to define an anytime-valid test. With this choice, the process $q_{(1)}, q_{(2)} \ldots$, is a typical instance of a *prequential plug-in* likelihood, that is often based on suitable smoothed likelihood-based estimators [11]. The phrase 'prequential' is due to [6], who reinvented and investigated the method, which is in fact much older, going back to [45]; see [29] for its history. The rationale behind this method is the following. Suppose the data are actually sampled from a distribution according to which the hazard ratio is θ. For sufficiently large initial risk sets, that is, if $\bar{y}_0^A$ and $\bar{y}_0^B$ are not too small, by the law of large numbers, the smoothed maximum likelihood estimate $\breve{\theta}_{(k)}$ will with high probability be close to θ. Therefore, $p_{\breve{\theta},(k)}$ will behave more and more like the real $p_{\theta,(k)}$ from which data are sampled. Thus, the process $S^{\text{preq}}_{\theta_0}$, will behave more and more similarly to the "correct" partial likelihood ratio (3.1).

Although we will not do so in the experiments to follow, in principle we could also extend this approach to incorporate prior knowledge or guesses about the hazard ratio under $\mathcal{H}_1$ by adding more virtual data points when calculating estimator $\breve{\theta}_{(k)}$. Viz. the remark underneath (3.5), this will not affect AV Type-I error validity. Technically, we need only modify the $p_{\theta,0}$ term in (3.7), which then plays a role analogous to a Bayesian prior: the more virtual data points we add in each group, the more influence they have on the likelihood; the ratio of virtual points in both groups determines our prior expectation of the hazard ratio under the alternative. As shown in Appendix A.4, we could also directly use a Bayesian prior, making our approach an instance of Robbins' [5] method of mixtures.

Example 3.2. If we set $\theta_0 = 1$ in (2.3), then the null expresses that both groups are indistinguishable, corresponding to a likelihood that, at each event time $T^{(k)}$, is proportional to the ratio between the number of participants still at risk in each group at that time. In a randomized clinical trial this would mean that the treatment has no effect. Interestingly and importantly, the anytime-valid Type-I error guarantee holds even if the alternative is arbitrarily misspecified—it may be the case that *if* the alternative is true, then the proportional hazard assumption does not
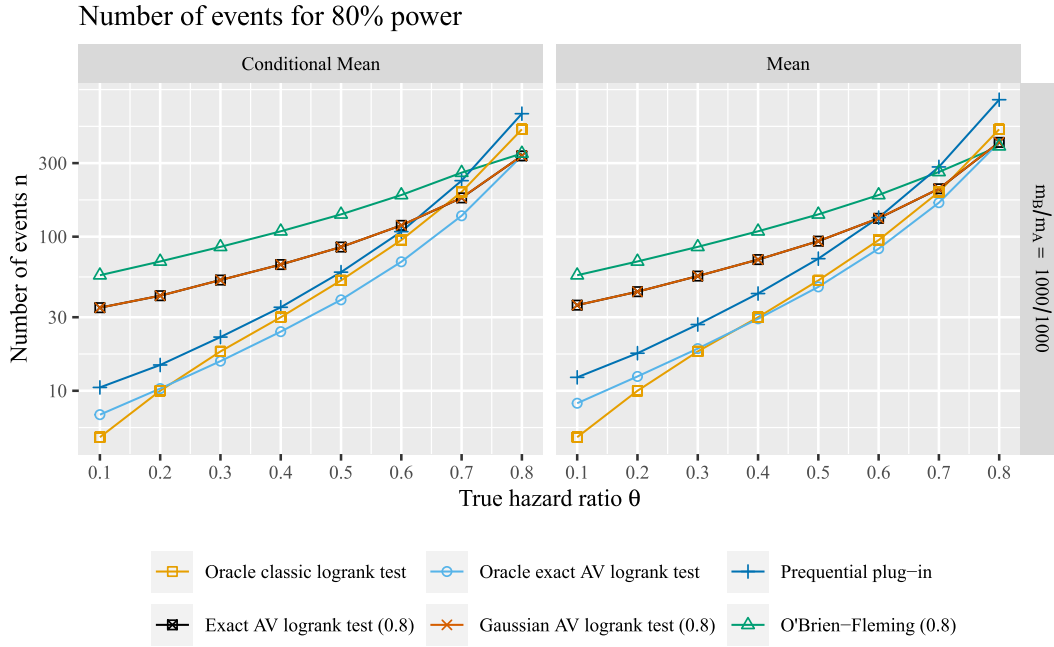
Number of events for 80% power

Figure 1: We show the number of events at which one can stop retaining 80% power at $\alpha = 0.05$ using the process $S_{\theta_0,t}^{\theta_1}$ with $\theta_0 = 1$ and $\theta_1 = 0.80$ when the true hazard ratio θ generating the data are different from θ_1. "Oracle" means that the method is specified with knowledge of the true θ, which in reality is unknown. Note that the y-axis is logarithmic.

hold. In particular, if we take $\mathcal{H}_1' : \theta \neq 1$ and use $S_{\theta_0,t}^{\text{preq}}$ as defined underneath (3.7), we can think of our test as really testing 'the treatment has no effect' against general alternative 'the treatment has an effect', where the test will have power as soon as there exists *some* $\theta \neq 1$ that will eventually lead to a higher partial likelihood than $\theta = 1$—this will often be the case even if the proportional hazards assumption is substantially violated. On the other hand, we can, of course, not make any claims about optimality in the GROW or any other sense in this misspecified case.

Now assume again that the model is well-specified (data are generated from a distribution satisfying proportional hazards for some θ). Then given a target value θ_1—a minimal clinically relevant effect size—the *worst-case* logarithmic growth rate of $S_{\theta_0,t}^{\text{preq}}$ will in general be smaller than that of the GROW $S_{\theta_0,t}^{\theta_1}$. Nevertheless, $S_{\theta_0,t}^{\text{preq}}$ can come close to the optimal for a whole range of potentially data-generating θ and may thus sometimes be preferable over choosing $S_{\theta_0,t}^{\theta_1}$. More precisely, the use of a prior allows us to exploit favorable situations in which θ is even smaller (more extreme) than θ_1. In such situations, the GROW $S_{\theta_0,t}^{\theta_1}$ is effectively misspecified. By using $S_{\theta_0,t}^{\text{preq}}$ that learns from the data, we may actually obtain a test martingale that grows faster than the GROW $S_{\theta_0,t}^{\theta_1}$, which is fully committed to detecting the worst-case θ_1.

In Figure 1, we illustrate such a situation where we start with 1000 participants in both groups (we defer general recommendations for when to use $S_{\theta_0,t}^{\theta_1}$ and when to use $S_{\theta_0,t}^{\text{preq}}$

to Section 8.4; for now, we merely aim to explain the difference). We generated data using different hazard ratios, and used a 'misspecified' $S_{\theta_0,t}^{\theta_1}$ that always used $\theta_1 = 0.8$. Note that while this is still the GROW (minimax optimal) test martingale for $\mathcal{H}_1 : \theta \leq \theta_1 \leq 0.8$. If we knew the true θ, we could use the test martingale $S_{\theta_0,t}^{\theta}$—it grows faster. We will call the test based on this latter martingale the *oracle* exact AV logrank test because it is based on inaccessible (oracle) knowledge. We estimated the number of events needed to reject the null with 80% power for $S_{\theta_0,t}^{0.8}$, the oracle $S_{\theta_0,t}^{\theta}$, and the prequential plug-in $S_{\theta_0,t}^{\text{preq.}}$. In all cases, we used the aggressive stopping rule that stops as soon as the statistic in question crosses the threshold $1/\alpha = 20$. We see that, as the true θ gets smaller than 0.8, we need fewer events using the GROW test $S_{\theta_0,t}^{0.8}$ (the data are favorable to us), but using the oracle exact AV logrank test we get a considerable additional reduction. The prequential plug-in $S_{\theta_0}^{\text{preq.}}$ 'tracks' the oracle $S_{\theta_0,t}^{\theta}$ by learning the true θ from the data: for θ near 0.8, it behaves worse (more data are needed) than $S_{\theta_0,t}^{0.8}$ (which knows the right θ from the start), but for $\theta < 0.6$ it starts to behave better. For comparison we also added the methods discussed in Section 4.1. Notably, the O'Brien-Fleming procedure, even though unsuitable for optional continuation, needs even more events than the misspecified AV logrank test $S_{\theta_0,t}^{0.8}$ as soon as θ goes below 0.8. The simulations were performed using exactly the same algorithms as for Figure 4 so the y-axis at $\theta = 0.8$ coincides with that of Figure 4, but now with absolute rather than relative numbers; details are described in Appendix A.5.

3.3 Tied Observations

Here, we propose a sequential test for applications where events are not monitored continuously, but only at certain observation times. In this case, more than one event may be witnessed in the time interval between two observation moments. Since the order in which these observations are made would be unknown, our previous approaches fail to offer a satisfactory sequential test. Assume that we make observations at times $t_0 < t_1 < t_2 < \cdots$ that are fixed before the start of the study. Even though we assume the absence of censoring in this section, this approach can be adapted to its presence under an additional common assumption: that the events reported between two observation times t_{k-1} and t_k precede any censorings, so that censored patients contribute fully to the risk sets under consideration. We assume that the available data are of the form $(O_1^A, O_1^B), (O_2^A, O_2^B), \ldots,$ where O_k^A and O_k^B are the number of events witnessed in each group in the time interval $(t_{k-1}, t_k]$, and $O_k = O_k^A + O_k^B$ is the total. Notice that since the observation times are discrete, we can index the observations by k instead of t_k. For each k, let $\bar{y}_k^A = \sum_{j \in A} y_{t_k}^j$ the number of participants at risk at time t_k, define similarly $\bar{y}_k^B$, and let $\bar{y}_k = \bar{y}_k^A + \bar{y}_k^B$ be the total. We derive an anytime-valid test—a test valid at any observation time—for the problem (2.3), where the hazard ratio under the null hypothesis is $\theta_0 = 1$, as in Example 3.2. The reason for this restriction in the null hypothesis—only $\theta_0 = 1$ is allowed—will soon become clear. Observe that, at time t_k, conditionally on $(\bar{y}_{k-1}^A, \bar{y}_{k-1}^B)$ and the total number of events O_k, the number of events O_k^B in group B follows a hypergeometric distribution. This implies that, conditionally on $(\bar{y}_{k-1}^A, \bar{y}_{k-1}^B, O_k)$, the conditional likelihood of observing O_k^B is $p_k(O_k^B) = p_{\text{Hyper}}(O_k^B; \bar{y}_{k-1}, \bar{y}_{k-1}^B, O_k)$, where p_{Hyper} is the probability mass function of a hypergeometric random variable, that is,

$$p_{\text{Hyper.}}(o^B; \bar{y}, \bar{y}^B, o) = \frac{\binom{\bar{y}^B}{o^B}\binom{\bar{y}-\bar{y}^B}{o-o^B}}{\binom{\bar{y}}{o}}.$$

With this observation at hand, we can build, analogously to (3.5) from the continuous-monitoring case, anytime-valid tests based on partial likelihood ratios,

$$S_{1,k}^q = \prod_{l \leq k} \frac{q_l(O_l^B)}{p_l(O_l^B)}, \tag{3.8}$$

where each q_k is a conditional distribution on the possible values of O_k^B that only depends on the data up to time t_{k-1}. Following the same steps as in Section 3, a sequential test based on monitoring whether $S_{1,k}^q$ crosses the threshold $1/\alpha$ is also anytime valid at level α.

Lemma 3.3. *Let $t_\kappa \in \{t_1, t_2, \ldots\}$ be an arbitrary random time. The test $\xi_{1,\kappa}^q$ given by $\xi_{1,\kappa}^q = \mathbf{1}\{S_{1,\kappa}^q \geq 1/\alpha\}$, where $S_{1,\kappa}^q$ is as in (3.8), has type-I error bounded by α, that is,*

$$\mathbf{P}_0\{\xi_{1,\kappa}^q = 1\} \leq \alpha,$$

under any distribution $\mathbf{P}_0$ such that the hazard ratio is $\theta = 1$.

Just as in the proof of Proposition 3.1, this lemma is shown by a combination of the martingale property of $S_{1,k}^{\theta_1}$ and Doob's maximal inequality. Therefore, we omit the proof of Lemma 3.3.

In order to obtain an optimal test under a particular hazard ratio θ_1—an alternative hypothesis—, it is necessary to compute the partial conditional likelihood for the data under the alternative of having observed O^B given $(\bar{y}_{k-1}^A, \bar{y}_{k-1}^B, \bar{N}_{k-1})$ (here $\bar{N}_k$ is the total number of observations up until time t_k). This conditional likelihood is given by Fisher's noncentral hypergeometric distribution with parameter ω. Unfortunately, ω depends on the baseline hazard function λ, which is assumed to be unknown (see Appendix A.3 for details). It is for this reason that we restrict the null hypothesis to $\theta_0 = 1$. Luckily, since the test based on $S_{\theta_0,t}^q$ remains valid even if q is only approximately correct, this problem can be skirted. As also noted by [21], when the times between observations are short, the parameter ω is well approximated by θ_1, the hazard ratio under the alternative hypothesis—no knowledge of λ is needed for the approximation. With this in mind, we put forward the use of $S_{\theta_0,k}^{\theta_1}$

$$S_{1,k}^{\theta_1} := \prod_{l \leq k} \frac{p_{\theta_1,k}(O_k^B)}{p_{1,k}(O_k^B)},$$

where $S_{1,k}^{\theta_1}$ is a an instance of (3.8) with $q_k(O_k^B) = p_{\theta_1,k}(O_k^B)$ and $p_{\theta_1,k}(O_k^B)$ is Fisher's noncentral hypergeometric distribution with parameter $\omega = \theta_1$, that is,

$$p_{\theta_1,k}(o^B) = p_{\text{FNCH}}(o^B; \bar{y}, \bar{y}^B, o, \omega = \theta_1)$$
$$= \frac{\binom{\bar{y}^B}{o^B}\binom{\bar{y}-\bar{y}^B}{o-o^B}\theta_1^{o^B}}{\sum_{\max\{0,o^B-\bar{y}^B\}\leq u \leq \min\{\bar{y}^B, o^B\}} \binom{\bar{y}^B}{u}\binom{\bar{y}-\bar{y}^B}{o^B-u}\theta_1^u}.$$

We remark that despite $p_{(\theta_1),k}$ being only approximately the correct distribution for the observations under the alternative, type-I error guarantees are not compromised (see the discussion on luckiness in Section 3.2). In any case, this approximation is accurate when the time between two consecutive observation times is not very long and when the number of tied observations is small. Two reassuring remarks are in order. First, in the special case when only one observation is made in each time interval between two consecutive observation moments, the statistic $S_{1,k}^{\theta_1}$ reduces to the continuously monitored AV logrank test (3.1) at time t_k. Second, the score test associated to $S_{1,k}^{\theta_1}$ coincides with the logrank test as is conventionally computed in the presence of ties.

4. A GAUSSIAN APPROXIMATION TO THE AV LOGRANK TEST

In this section we present an approximation to the AV logrank test introduced in the previous section. This is based

on a Gaussian approximation to the logrank statistic. The approximation is of interest for two reasons. First, in practical situations, only the logrank Z-statistic (a standardized form of the classic logrank statistic) and other summary statistics may be available—and not the full risk-set process. This is often the case in medical meta-analyses, where the full data sets (Individual Participant Data, IPD, for each trial) are often not available, but summary statistics are. The Gaussian approximation to the AV logrank statistic can then often still be used (we remark though that in the BCG trial meta-analysis [33] referred to in the Introduction, there was IPD access and the exact anytime-valid logrank test was performed). The second reason, which we address in Section 4.1, is related to the fact that α-spending and group-sequential approaches, which we use as benchmarks, are also based on Gaussian approximations to the classic logrank statistic. Consequently, the behavior of the Gaussian approximation gives further insights into how the AV logrank statistic compares to group-sequential and α-spending approaches as well. We henceforth focus on $\theta_0 = 1$, a central case of interest also considered in Example 3.2.

Our general strategy is close in spirit to that followed in the construction of the exact AV logrank statistic in Section 3. We build likelihood ratios using a classic Gaussian approximation for the distribution of the original logrank statistic [32]. If the distribution of this statistic was exactly normal, we could monitor continuously its likelihood ratio. We show through extensive simulation in which regimes this approximation behaves similarly to the AV logrank statistic.

We begin by recalling the definition of the Z-score associated to the classic logrank test. Let $E_i^B = O_i p_i^B$ with $p_i^B = \bar{y}_i^B / (\bar{y}_i^A + \bar{y}_i^B)$ be the expected (under the null) number of events witnessed in the time interval $(t_{i-1}, t_i]$ in group B, and let $V_i^B = O_i\, p_i^B (1 - p_i^B) \frac{\bar{y}_i - O_i}{\bar{y}_i - 1}$ be its variance. After k observations the Z-score associated to the classic logrank statistic, Z_k, is given by

$$Z_k = \frac{\sum_{i \le k} \left\{ O_i^B - E_i^B \right\}}{\sqrt{\sum_{i \le k} V_i^B}}. \tag{4.1}$$

The numerator in the definition of Z_k is the classic logrank statistic $H_k = \sum_{i \le k} \left\{ O_i^B - E_i^B \right\}$, which is typically interpreted as the cumulative difference between observed counts O_i^B and the expected counts E_i^B in Group B. The factor $\frac{\bar{y}_i - O_i}{\bar{y}_i - 1}$ found in V_i^B can be interpreted as a multiplicity correction, that is, a correction for ties [15, p. 207]. When only one event is witnessed between two consecutive observation times, then $O_i = 1$, $E_i^B = p_i^B$, and $V_i^B = p_i^B (1 - p_i^B)$. We remark that the above formulation is also found in the work of Cox [3, (26)].

Fix some initial m^A, m^B and let $\gamma = m^B / m^A$. Schoenfeld [32] showed that in the limit for $m^A \to \infty$, $m^B = \gamma m^A$, $k \to \infty$, $k/m^A \to 0$, and $\log \theta = O((m^A + m^B)^{-1/2})$, the distribution of Z_k under any distribution with hazard ratio θ_1 converges to the distribution of a normal with unit variance and mean

$$\mu_1 = \log(\theta_1) \sqrt{m^B m^A / (m^A + m^B)^2}. \tag{4.2}$$

Appendix C provides more details about this result. If Z_k were exactly $N(\mu_1, 1)$ distributed, then an easy calculation shows that the logrank partial likelihood ratio statistic $S_{1,k}^{\theta_1}$ would be equal to a likelihood ratio of two Gaussians, given by

$$S_k^G := \exp\left(-\frac{1}{2}\bar{N}_k \mu_1^2 + \sqrt{\bar{N}_k}\mu_1 Z_k \right), \tag{4.3}$$

where $\bar{N}_k$ is the total number of observations up until time t_k and μ_1 is given by (4.2). We thus put forward S_k^G as a Gaussian approximation to the logrank statistic $S_{1,k}^{\theta_1}$. For an arbitrary random observation time $t_K \in \{t_1, t_2, \dots\}$, we refer to the test $\xi_K^G = \mathbf{1}\{S_K^G \ge 1/\alpha\}$ as the Gaussian AV logrank test for (2.3). Recall that we test $\theta_0 = 1$, which corresponds to the asymptotic mean of the Z-score under the null hypothesis being $\mu_0 = 0$. Clearly, the approximation may fail in practice for at least two reasons: first, m^A, m^B, k and θ_1 may not fall under Schoenfeld's asymptotic regime; second, Schoenfeld's result has been derived in a nonsequential setting, whereas we aim to use it in a sequential setting with optional stopping. Thus, in Appendix C.1 extensive simulations are performed to show in which regimes the Gaussian logrank test retains type-I error guarantees. In Appendix C.2, it is shown that, under continuous monitoring, the Gaussian AV logrank test tends to be more conservative—it needs more data than the exact one. The conclusion is the following: S_K^G can be used for designs with balanced allocation, and it approximates $S_{1,K}^{\theta_1}$ well for hazard ratios between 0.5 and 2.

We now compare the rejection regions defined by the Gaussian logrank test to those of continuously monitoring using α-spending and group-sequential approaches.

4.1 Rejection Region and α-Spending

In this section we compare the rejection regions of the Z-scores for which α-spending approaches and the AV logrank test for the null hypothesis of no effect (hazard ratio $\theta_0 = 1$, Example 3.2). The two main α-spending approaches discussed here are due to Pocock [25] and O'Brien and Fleming [22]. We provide two reasons why the main focus of the comparison, however, will be on the O'Brien-Fleming approach. Firstly, in retrospect, Pocock himself believes that his approach leads to boundaries that are unsuitable [26]. One main feature of the Pocock procedure is that the rejection regions are the same regardless of whether the (interim) analyses are conducted at the start or the end of the trial. In practice this leads to many stopped trials for benefits based on (too) small sample sizes and with unrealistically large treatments effects [26]. In contrast, the rejection boundary of the O'Brien-Fleming is more conservative at the start than at the end of the trial. Secondly, the Pocock procedure

only allows for a finite number of planned analyses and, therefore, cannot be monitored continuously, whereas this is possible with the O'Brien-Fleming α-spending approach. Hence, the fair comparison is between the two procedures (the AV logrank test and the O'Brien-Fleming α-spending approach) that allow for continuous monitoring.

We begin by specifying the rejection regions for both the Gaussian AV logrank test and that of the O'Brien-Fleming α-spending procedure. For the Gaussian AV logrank we compute the region for the Z-score that rejects the null hypothesis. Indeed, using (4.3), we can compute that whenever $m^A = m^B$, the null hypothesis is rejected as soon as

$$Z_n \geq \frac{\sqrt{n}}{4}\ln(\theta_1) - \frac{2}{\sqrt{n}}\frac{\log(\alpha)}{\log(\theta_1)} \quad \text{if} \quad \theta_1 > 1, \text{ or}$$

$$Z_n \leq \frac{\sqrt{n}}{4}\ln(\theta_1) - \frac{2}{\sqrt{n}}\frac{\log(\alpha)}{\log(\theta_1)} \quad \text{if} \quad \theta_1 < 1.$$

The O'Brien-Fleming procedure is based on a Brownian-motion approximation to the sequentially computed logrank statistic Z-score. Indeed, for large values of $n_{\max}$ and $t \in [0,1]$, the process $t \mapsto \sqrt{\frac{\lfloor tn_{\max}\rfloor}{n_{\max}}}Z_{\lfloor tn_{\max}\rfloor}$ can be approximated by a Brownian motion B_t. We stress the fact that $n_{\max}$ has to be set in advance.[1] If B_t is a Brownian motion, the reflection principle, a well-known but nontrivial application of the symmetry of B_t, implies that

$$\mathbf{P}\{\max_{0 \leq t \leq 1} B_t \geq c\} = 2\mathbf{P}\{B_1 \geq c\}$$

Since B_1 is Gaussian with mean zero and standard deviation 1, setting $c = q_{1-\alpha/2}$, the $(1-\alpha/2)$-quantile of a standard Gaussian distribution, then

$$\mathbf{P}\{\max_{0 \leq t \leq 1} B_t \geq q_{1-\alpha/2}\} = \alpha.$$

This implies that

$$\mathbf{P}\{\max_{n \leq n_{\max}} \sqrt{n}Z_n \geq \sqrt{n_{\max}}q_{1-\alpha/2}\} \approx \alpha,$$

or, in other words, the procedure that continuously monitors whether the Z-score crosses the boundary $\sqrt{n_{\max}}q_{1-\alpha/2}$ guarantees approximate type-I error α. Given a hazard ratio θ_1 under the alternative hypothesis, $n_{\max}$ can be set to achieve a desired type-II error. The left-handed procedure can be worked out similarly, and we obtain that, for $m^A = m^B$, the continuous-monitoring version of the O'Brien-Fleming procedure rejects as soon as

$$Z_n \geq \sqrt{\frac{n}{n_{\max}}}q_{1-\alpha/2} \quad \text{if} \quad \theta_1 > 1 \text{ (right-sided test), or}$$

---

[1] One might think that a maximum sample size is implicit also in the AV logrank test: if we start with $m = m^A + m^B$ participants, then there can be no more observations after m events have taken place. However, in many practical applications, this may take exceedingly long (or may never even happen), and then the α-spending approaches discussed here would have to set an $n_{\max} \ll m$. Once this $n_{\max}$ has been chosen, one cannot go beyond it, whereas with the AV logrank test, there is no such restriction.

$$Z_n \leq \sqrt{\frac{n}{n_{\max}}}q_{1-\alpha/2} \quad \text{if} \quad \theta_1 < 1 \text{ (left-sided test).}$$

The two regions of the Z-statistic values share an important feature: they are more conservative to reject the null hypothesis at small sample sizes than at larger ones, requiring more extreme values for the Z-statistic at the start of the trial. This sets them apart from the Pocock spending function that requires equally extreme values for the Z-statistic at small and large sample size. Figure 2 shows both the Gaussian AV logrank and the O'Brien-Fleming α-spending rejection regions. Additionally, Figure 2 shows the boundary of the Pocock α-spending function for 10 interim analyses. Note that the definition of the AV logrank test rejection region requires a very explicit value for the effect size $\theta_1 = \theta_{\min}$ of minimum clinical relevance, while that value is implicit in the definition of the α-spending rejection region: To specify an maximum sample size $n_{\max}$ to achieve a certain power, an effect size of minimal interest is also assumed. A fixed-sample-size analysis designed to detect a minimum hazard ratio of 0.7 would need 195 events to achieve 80% power if the true hazard ratio is also 0.7. A sequential analysis using α-spending requires a slightly larger maximum number of events: 205 with the O'Brien-Fleming spending function; 245, with the Pocock α-spending function—when we design for 10 interim analyses. We investigate the number of events needed by the Gaussian AV logrank test in Appendix C.2. For the α-spending procedures continuing beyond $n_{\max}$ is problematic. This is not the case for the AV logrank test, as it allows for unlimited monitoring, then $n_{\max}$ is only a soft constraint on the study—there is no penalty in type-I error for continuing after $n_{\max}$ events have been witnessed.

The benefit of a sequential approach is that if there is evidence that the hazard ratio is more extreme than it was anticipated under the alternative hypothesis, we can detect that with fewer events than the maximum sample size. The left column of Figure 3 illustrates that we benefit because the true hazard ratio could be more extreme than we designed for (e.g. 0.5 instead of 0.7; a larger risk reduction in the treatment group) and the data reflects that. We also benefit from a sequential analysis if the true hazard ratio is 0.7 but by chance the values of our Z-statistics are more extreme than expected. The major difference between α-spending approaches and the AV logrank test is that the AV test does not require to set a maximum sample size. It in fact allows to indefinitely increase the sample size without ever spending all α. An α-spending approach designed to have 80% power will miss out on rejecting the null hypothesis in 20% (the type-II error) of the cases as is illustrate in the bottom middle plot of Figure 3 by the sample paths that remain (dark) green. In contrast, the AV logrank test can potentially reject with 100% power by continue sampling. In the sample paths of 500 events in Figure 3, all but one sample path of Z-statistics could be rejected at a larger
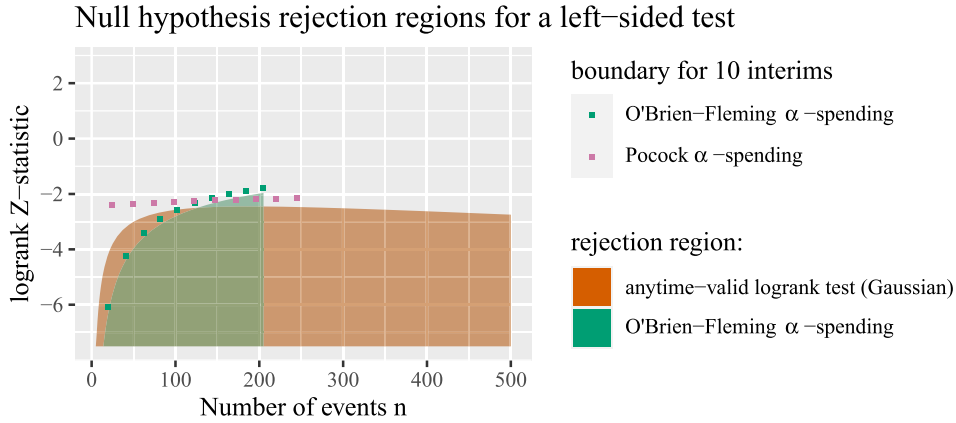
Null hypothesis rejection regions for a left−sided test

Figure 2: Left-sided rejection regions for continuous-monitoring using O'Brien-Fleming α-spending or the Gaussian AV logrank test. Allocation is balanced ($m^A = m^B$) and $\alpha = 0.05$. Also shown are the O'Brien-Fleming and Pocock α-spending boundaries for 10 interim analyses. The α-spending boundaries are designed to have 80% power when detecting a hazard ratio 0.7. For more details, including the values of $n_{\max}$, see Section 4.1.

sample size by the AV logrank test. By extending the trial, the AV logrank test can potentially have 100% power if the true hazard ratio is at least as small as the hazard ratio set for minimum clinical relevance in the design of the test. Still, type-I error is controlled. The bottom right plot of Figure 3 shows two null sample paths with a true hazard ratio of 1 that are rejected by the O'Brien-Fleming α-spending region, but not by the AV logrank test. Here, the AV logrank test is more conservative.

It is known that α-spending methods behave poorly in case of unbalanced allocation [51]. In Appendix C.1 we showed that our Gaussian approximation to the logrank test is also not an E-variable in case of unbalanced allocation. Our exact AV logrank test, however, is an E-variable under any allocation since it is defined directly on the risk-set process (3.3). This suggests that if the complete data set is available and allocation is unbalanced, the exact logrank test should be preferred over the Gaussian approximation and the α-spending methods.

5. OPTIONAL CONTINUATION AND LIVE META-ANALYSIS

In this section, we address optional continuation and live meta-analysis—the continuous aggregation of evidence from multiple experiments. For instance, data could come from medical trials conducted in different hospitals or in different countries. In such cases, we compare a global null hypothesis $\mathcal{H}_0$ that is addressed in all trials (for instance, $\theta_0 = 1$) to an alternative hypothesis $\mathcal{H}_1$ that allows for different hazard ratios in each experiment. The present approach covers even the case in which the decision to start each experiment might depend on the observations made in experiments that are already in progress. Assume that there are k_E experiments, $E_{(1)}, \ldots, E_{(k_E)}$, ordered by their respective

starting times $V_{(1)} \leq \cdots \leq V_{(k_E)}$, each performed on different and independent populations. Assume further that the starting time $V_{(k)}$, of experiment $E_{(k)}$ depends only on the data observed in the ongoing experiments $E_{(1)}, \ldots, E_{(k-1)}$. If each experiment $E_{(k)}$ monitors the AV logrank statistic $S_{\theta_0, t}^k$, where $S_{\theta_0, t}^k = 1$ for $t \leq V_{(k)}$, then the product statistic $S_{\theta_0, t}^{\mathrm{meta}} = \prod_{i \leq k_E} S_{\theta_0, t}^i$ is a test martingale with respect to the filtration generated by all observations. Consequently, the meta-test based on it enjoys anytime validity.

Proposition 5.1. *Let τ be any random time. The test $\xi_{\theta_0, \tau}^{\mathrm{meta}}$ given by $\mathbf{1}\{S_{\theta_0, \tau}^{\mathrm{meta}} \geq 1/\alpha\}$, where $S_{\theta_0, t}^{\mathrm{meta}} = \prod_{i \leq k_E} S_{\theta_0, t}^i$, has type-I error smaller than α.*

This result follows from a reduction to independent left-truncation—we refer to left-truncation in the specific sense defined by Andersen et al. [1]. Indeed, even in the presence of dependencies on other studies, the observations made in $E_{(k)}$ can be regarded as a left-truncated sample. Here, the time at which observation in $E_{(k)}$ is started is random and only participants that have not witnessed an event are recruited into the study. One may worry that these dependencies may alter the sequential properties of $S_{\theta_0, t}^{\mathrm{meta}}$, but this is not the case. Since the truncation time for $E_{(k)}$ is based on data that are independent of that of experiment $E_{(k)}$—it is possibly based on the observations made in all other experiments, it follows from results of Andersen et al. [1] (see Appendix A.2) that the sequential-multinomial interpretation of the partial likelihood for the truncated data remains valid. Consequently, so does the sequentially computed AV logrank statistic and the product statistic $S_{\theta_0, t}^{\mathrm{meta}}$. By continuously monitoring $S_{\theta_0, t}^{\mathrm{meta}}$, we effectively perform an *online, cumulative* and possibly *live* meta-analysis that remains valid irrespective of the order in which the events of the different trials are observed. Importantly, unlike in α-spending approaches, the maximum number of trials and
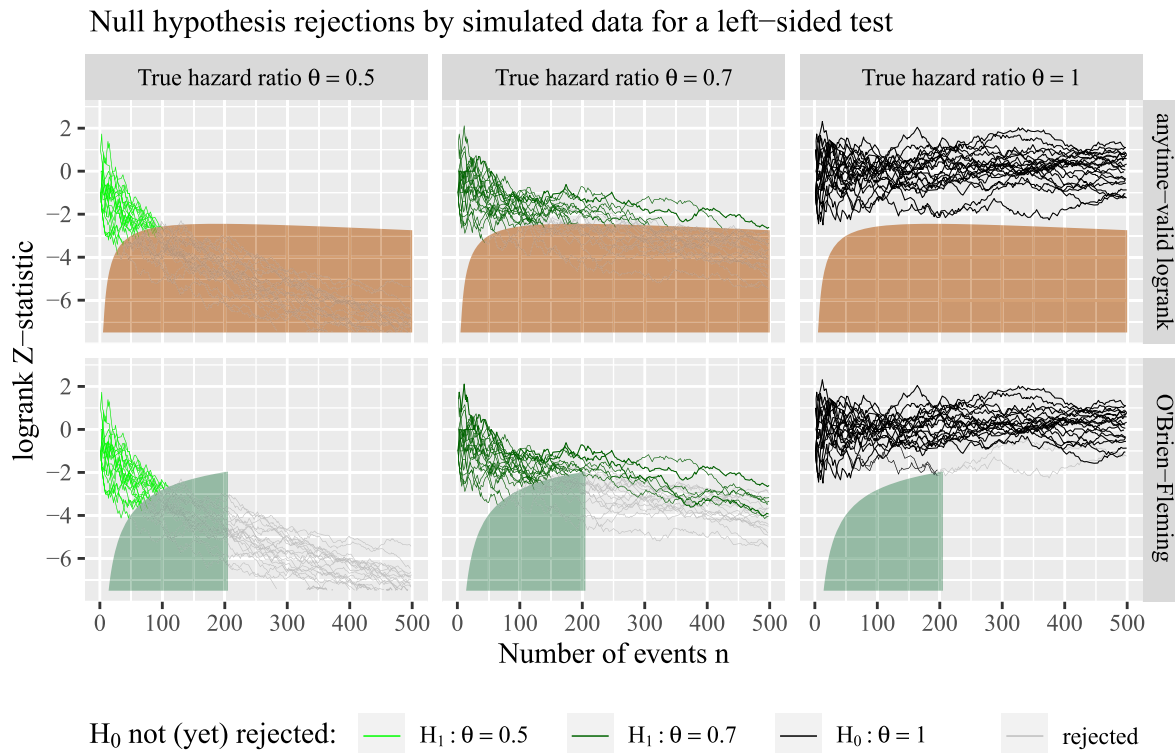
Figure 3: Null hypothesis rejections on simulated data. The rejection regions are the same as shown in Figure 2 (designed to detect a hazard ratio of 0.7 with 80% power). Data are simulated under balanced allocation ($m_1 = m_0 = 5000$) and as time-to-event data with possible ties. The logrank Z-statistic does not have a value for all n; it sometimes jumps with several additional events at a time.

6. ANYTIME-VALID CONFIDENCE SEQUENCES

Anytime-valid (AV) confidence sequences correspond to anytime-valid tests in the same way fixed-sample tests correspond to confidence sets. Indeed, it is possible to "invert" a fixed-sample test to build a confidence set: the parameters of the null hypothesis that are not rejected by a the test form a confidence set. Analogously, test martingales can be used to derive AV confidence sequences [5, 16, 12, 13]. In our setting, a $(1 - \alpha)$-AV confidence sequence is a sequence of confidence sets $\{CI_t\}_{t \geq 0}$, such that

$$\mathbf{P}_\theta\{\theta \notin CI_t \quad \text{for some} \quad t \geq 0\} \leq \alpha. \quad (6.1)$$

We call it an AV confidence *interval* if each confidence set CI_t is itself an interval. A standard way to design $(1 - \alpha)$-AV confidence intervals, translated to our logrank setting, is to use a prequential plug-in test martingale $S_{\theta_0, t}^{\text{preq}}$ as in

Section 3.2 (or the Bayesian alternative discussed in Appendix A.4). At time t, one reports $CI_t = [\theta_t^L, \theta_t^U]$ where CI_t is the smallest interval containing the values of θ_0 such that $S_{\theta_0, t}^{\text{preq}} > 1/\alpha$ outside this interval. Ville's inequality readily implies that this is indeed an AV confidence interval. The same construction can be made for arbitrary instances of $S_{\theta_0, t}^q$ as in (3.5).

7. POWER AND SAMPLE SIZE

In this section, we investigate the power properties of the AV logrank test—we will study specific stopping times. We have seen that by observing arbitrarily long sequences of events the logrank test can achieve type-II errors that are as close to zero as desired. However, in practice it is necessary to plan for a maximum number of events $n_{\max}$ so that either the experiment is stopped as soon as the null hypothesis is rejected or when $n_{\max}$ events have been observed. In the latter case, there is no evidence to reject the null hypothesis. We assess via simulation the value of $n_{\max}$ needed to guarantee 20% type-II error (80% power) for the exact and Gaussian AV logrank tests. We compare this to the $n_{\max}$ needed to achieve the same power using the continuous-monitoring O'Brien-Fleming α-spending procedure introduced in the previous section, and the fixed-sample-size classic logrank
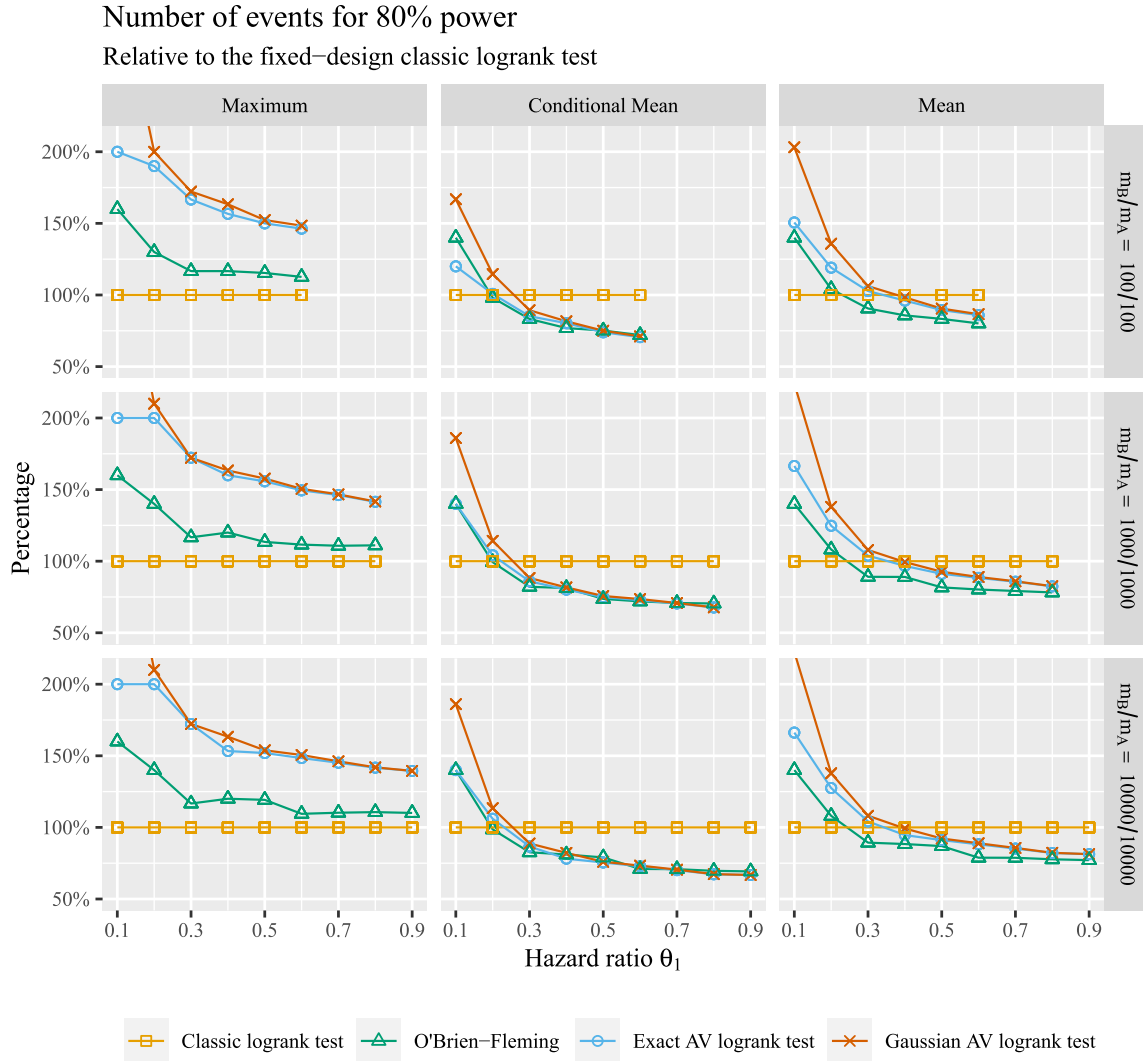
Number of events for 80% power

Relative to the fixed−design classic logrank test

Figure 4: Maximum, expected (Mean) number of events needed to reject the null hypothesis with 80% power. 'Conditional Mean' makes reference to the number of events needed given that the null hypothesis is indeed rejected. The maximum number of events needed using AV logrank statistics is higher than that of a fixed-sample test, but lower in expectation (see Section 7). All simulations are performed with $\alpha = 0.05$ and tests are designed to detect the hazard ratio θ_1 shown on the x-axis. Data are generated using that same hazard ratio. The classical logrank test needs the following sample sizes (number of events) $n(\theta_1)$ for an 80%-power design to detect hazard ratio θ_1: $n(0.1) = 5$, $n(0.2) = 10$, $n(0.3) = 18$, $n(0.4) = 30$, $n(0.5) = 52$, $n(0.6) = 95$, $n(0.7) = 195$, $n(0.8) = 497$ and $n(0.9) = 2228$. These sample sizes represent the 100% line in all plots.

test. Figure 4 show simulation results establishing three types of sample sizes. The leftmost panels ("Maximum") shows the sample size $n_{\max}$ described earlier, which would be required to design the experiment. We stress the fact that using the classic logrank test or α-spending designs events beyond $n_{\max}$ cannot be analyzed. The rightmost panel of Figure 4 ("Mean") shows the sample sizes that capture the expected duration of the trial. It expresses the mean number of events, under the alternative hypothesis, that will be observed before the trial can be stopped. Here, for the AV logrank tests, we use the aggressive stopping rule that stops

as soon as $S^{\theta_1}_{\theta_0, t} \geq 1/\alpha = 20$ or $n = n_{\max}$. In case of α-spending approaches and the AV logrank test this number of events is always smaller than the maximum needed in the design stage. Lastly, the middle panel ("Conditional Mean") shows an even smaller number for those tests that have a flexible sample size: the expected stopping time *given* that the trial is stopped before the maximum $n_{\max}$ was reached— this only happens if the null is rejected. For comparison purposes, all sample sizes are shown relative to (i.e., divided by) the fixed sample size needed by the classical logrank test to obtain 80% power. Note that for small sample size

(for small hazard ratios), both the classic logrank test and O'Brien-Fleming α-spending are not recommended due to lack of type-I error control. They are based on Schoenfeld's Gaussian approximation, which underestimates the number of events required for hazard ratios far away from 1. For example, simulations show that for $\theta_1 = 0.1$, $n = 6$ or 7 events will be necessary—for small sample sizes the classical logrank test is not recommended due to lack of type-I error control. We give further details in Appendix A.5 (see also Figure 4). In summary, at all hazard ratios at which the Gaussian approximation to the classic logrank test is accurate (say for $\theta_1 \geq 0.3$), the mean number of events needed by the AV logrank tests is about the same or noticeably smaller than that needed when using a fixed-sample-size analysis.

8. DISCUSSION, FUTURE WORK AND RECOMMENDATIONS FOR PRACTICE

We introduced the AV logrank test, a version of the logrank test that retains type-I error guarantees under optional stopping and continuation. Extensive simulations reveal that, if we do engage in optional stopping, it is competitive with the classic logrank test (which neither allows in-trial optional stopping nor optional continuation) and α-spending procedures (which allow forms of optional stopping but not optional continuation). We provided an approximate (Gaussian) test for applications in which only summary statistics are available and also showed how the AV logrank test can be used in combination with prequential learning approaches, when no effect size of minimal clinical relevance can be specified. The GROW AV logrank tests (exact and Gaussian) are available in our `safestats` R package [43].

We end this paper by discussing how we avoid the issue of so-called *doomed trials* (Section 8.3) and providing a practical guideline for what method to use in what situation (Section 8.4). But first we discuss two highly important potential extensions: *including covariates other than group membership* and *staggered entries*.

8.1 Covariates: The Full Cox Model

From a purely theoretical perspective, it is straightforward to extend the AV logrank test to the situation when time-dependent covariates are present, making the underlying model equivalent to the full Cox proportional hazards model. We sketch how to do this extending the notation of Section 3. Assume therefore the presence of d covariates and let, for each participant i, $\mathbf{z}_t^i = (z_{t,1}^i, \ldots, z_{t,d}^i)$ be the covariate vector consisting of left-continuous time-dependent covariates $z_{t,1}^i, \ldots, z_{t,d}^i$. Denote by $\mathbf{z}_{(k)}^i$ the value of the covariates of participant i at the time $T_{(k)}$ when the kth event is witnessed. We let random variable $I_{(k)}$ denote the index of the patient to which the kth event happens, and consider the extended process $I_{(1)}, I_{(2)}, \ldots$ where the information that is available at time $T_{(k)}$ is, $I_{(1)}, I_{(2)}, \ldots, I_{(k)}$, and $z_{(1)}, \ldots, z_{(k)}$.

The conditional partial likelihood is now denoted $\mathbf{P}_{\beta,\theta}$ with $\theta > 0$, $\boldsymbol{\beta} \in \mathbb{R}^d$, and $\beta_\theta = \ln \theta \in \mathbb{R}$, defined as follows:

$$
\begin{aligned}
&\mathbf{P}_{\boldsymbol{\beta},\theta}\{I_{(k)} = i \mid \mathbf{z}_{(l)}^j, y_{(l)}^j;\ j = 1, \ldots, n;\ l = 1, \ldots, k\} \\
&\quad := P_{\beta,\theta}\{I_{(k)=i} \mid \mathbf{z}_{(k)}^j, y_{(k)}^j;\ j = 1, \ldots, n\}, \\
&\mathbf{P}_{\boldsymbol{\beta},\theta}\{I_{(k)} = i \mid \mathbf{z}_{(k)}^j, y_{(k)}^j;\ j = 1, \ldots, n\} \\
&\quad := p_{\boldsymbol{\beta},\theta,(k)}(\ i\ ) := \frac{\exp(\langle \boldsymbol{\beta}, \mathbf{z}_{(k)}^i \rangle + g^i \beta_\theta)}{\sum_{j \in \mathcal{R}_{(k)}} \exp(\langle \boldsymbol{\beta}, \mathbf{z}_{(k)}^j \rangle + g^j \beta_\theta)}.
\end{aligned}
\tag{8.1}
$$

This is consistent with Cox' [3] proportional hazards regression model: the probability that the ith participant witnesses 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 $\boldsymbol{\beta} = 0$, this is easily seen to coincide with the definition of $\mathbf{P}_\theta$ via (2.5) with $\theta = e^{\beta_\theta}$.

For simplicity we focus on the central case that $\mathcal{H}_0$ expresses that there is no treatment effect, i.e. we have composite null $\mathcal{H}_0 : \boldsymbol{\beta} \in \mathbb{R}^d; \theta = 1$. First we simplify even further and consider a simple alternative $\mathcal{H}_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1, \theta = \theta_1$ for a specific fixed vector $\boldsymbol{\beta}_1 \in \mathbb{R}^d$ and $\theta \neq 1$, with partial likelihoods given as above. We now want to extend (3.1) so that the corresponding test (3.2) is anytime-valid under all elements of $\mathcal{H}_0$, i.e. the appropriate extension of Proposition 3.1 holds (technically, for this we need the partial likelihood ratio to define an e-process). The factors in the numerator in (3.1) may simply be replaced by the corresponding partial likelihood factors $p_{\boldsymbol{\beta}_1,\theta_1,(k)}(\ i\ )$. Since the null is now composite, it is not immediately clear what to use for the denominator. However, there exist two (by now standard) methods to set the numerator such that the ratio does become an e-process [29], the *Universal Inference* (UI) method [47] and *Reverse Information Projection* (RIPr, [9]). In the present setting, UI is much more straightforward to use: it amounts to set the denominator, at each time t, equal to the running-maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{(k_t)}$ based on the k_t events that happened before time t, i.e. for given t, k_t is the largest k such that $T^{(k)} < t$. The partial likelihood ratio (3.1) then becomes

$$
\prod_{k:T^{(k)} \leq t} \frac{p_{\boldsymbol{\beta}_1,\theta_1,(k)}(I_{(k)})}{p_{\hat{\boldsymbol{\beta}}_{(k_t)},1,(k)}(I_{(k)})}.
\tag{8.2}
$$

We can now easily extend this approach to deal with the full $\mathcal{H}_1 : \theta \neq 1$ (testing whether there is any treatment effect at all, see Example 3.2) by learning $(\boldsymbol{\beta}_1, \theta_1)$ using a prequential approach as in Section 3.2. It is straightforward to extend Proposition 3.1 to show that with this choice, (8.2) provides an anytime-valid test. The advantage of this UI method is that it can be readily implemented by small modifications of existing software for finding the MLE in the Cox model; the disadvantage is that we may expect it to have suboptimal power unless the number of covariates is small compared to

the number of observed events [40]. Alternatively, we may employ a sequential version of the RIPr method. This will also lead to an anytime-valid test, but, at least currently, we do not know how to calculate it efficiently; for completeness, we describe this approach in Appendix B, where we also provide additional details about covariates in combination with ties. In future work, we aim to investigate and compare the UI and the RIPr approach in detail.

8.2 Staggered Entry

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, say, a two-day-after-surgery event changes when new participants enter and survive two days). Sellke and Siegmund [36] and Slud [38] show that, in general, martingale properties cannot be preserved under such staggered entry settings, but that asymptotic results are hopeful [36] as long as certain scenarios are excluded [38]. When all participants' risk is on the same (calendar) time scale (e.g., infection risk in a pandemic; staggered entry now amounts to left-truncation, which we can deal with), or new patients enter in large groups (allowing us to stratify), staggered entry poses no problem for our methods. But research is still ongoing into those scenarios in which our inference is fully AV for patient time under staggered entry, and those that need extra care.

8.3 Your Trial Is Not Doomed

In their summary of conditional power approaches in sequential analysis Proschan, Lan, and Wittes [27] write that low conditional power makes a trial futile. Continuing a trial in such case could only be worth the effort to rule out an effect of clinical relevance, when the effect can be estimated with enough precision. However, if "both conditional and revised unconditional power are low, the trial is doomed because a null result is both likely and uninformative" [27, p. 63]. While this is the case for all existing sequential approaches that set a maximum sample size, this is not the case for AV tests. Any trial can be extended and possibly achieve 100% power or in an anytime-valid confidence sequence show that the effect is too small to be of interest. This is especially useful for time-to-event data when sample size can increase by extending the follow-up time of the trial, without recruiting more participants. Moreover, new participants can always be enrolled either within the same trial or by spurring new trials that can be combined indefinitely in a cumulative meta-analysis.

8.4 Recommendations for Practical Use

Throughout this paper, we encountered different instantiations of the AV logrank test. Specifically, for alternatives as introduced in Section 2 that are 'disconnected' from $\mathcal{H}_0$, such as in (2.3) (for left-sided alternatives) or (3.6) (for two-sided alternatives) we presented partial likelihood ratios $S_{\theta_0,t}^{\theta_1}$ with the numerator based on a point alternative θ_1 as in (3.1) (one-sided case) or a mixture of two point alternatives (below (3.6)). For the full alternative $\mathcal{H}_1 : \theta \neq \theta_0$ we presented partial likelihood ratios $S_{\theta_0,t}^{\mathrm{preq}}$ with the numerator based on prequential plug-in likelihood as in Section 3.2. However, all these versions continue to provide anytime-valid Type-I error bounds under $\mathcal{H}_0$, no matter what the potential 'real' alternatives we may encounter; they may not even satisfy proportional hazards (Example 3.2). So, what instantiation of the test should we choose in practice?

This depends on the situation. If we have a clearly defined one-sided minimum clinically relevant effect and we are convinced that the proportional hazards assumption holds, then this provides a strong motivation for using the point alternative version $S_{\theta_0,t}^{\theta_1}$ since it has GROW status (Section 3.1): by Breiman's result detailed in Appendix A.1 it will give us the smallest expected stopping time in the worst-case over all alternatives. In contrast, if we are interested in whether there is treatment effect at all (i.e. $\mathcal{H}_1 : \theta \neq \theta_0$), but also if misspecification might be an issue (proportional hazards might be substantially violated, and then 'effect size' is not really well-defined) we should use the prequential version $S_{\theta_0,t}^{\mathrm{preq}}$, learning an alternative from the data. Yet, even *if* we have a minimally clinically relevant effect size in mind and we may assume proportional hazards, there are cases in which we advocate to use the prequential version after all, for the reason that, by varying θ_0 in $S_{\theta_0,t}^{\mathrm{preq}}$ via the construction in Section 6, this provides us with anytime-valid confidence sequences with width that is guaranteed to shrink toward the actual effect size—in contrast, it is easily seen that if we vary θ_0 in $S_{\theta_0,t}^{\theta_1}$ then the resulting AV confidence sequences may remain wide forever. Well-behaved (i.e., shrinking) AV confidence sequences can be useful or even essential if we aim to communicate the results (confidence sequences are much more intuitive to practitioners than accept/reject decisions accompanied by partial likelihood ratios). Also, in some cases it cannot be ruled out that, and it would be really useful to know whether, the actual effect is much stronger than the minimum clinically relevant one. A case in point are the covid vaccination trials in which the original test had as alternative $\mathcal{H}_1$ that vaccine efficacy was above 50%, which was viewed as the minimum clinically relevant value, but the data suggested it was really more than 80%, which changed the outlook altogether (for example, suggesting vaccination campaigns for the entire adult population rather than just the elderly). A test based on the point alternative $S_{\theta_0,t}^{\theta_1}$ with θ_1 corresponding to 50% could not easily have been transformed post-hoc into a test for testing whether the efficacy is larger than 80%, whereas with an AV confidence sequence based on $S_{\theta_0,t}^{\mathrm{preq}}$ this can be established right away [35]. Finally, we note that we only recommend use of the *Gaussian* AV log rank test if one really has to, for example, as in the situation described in Section 4, when only logrank Z-statistics are available; and even then, only in the regime

where it provides anytime-valid Type-I error guarantees, as explored in that section.

APPENDIX A. OMITTED PROOFS AND DETAILS

In this section we provide proofs and remarks omitted from previous sections. In Appendix A.1 we relate growth-rate optimality to the minimum expected stopping time. In Appendix A.2, we show that the AV logrank statistic is a continuous-time martingale, and show that this is also true for general patterns of incomplete observation, such as left truncation and filtering as a consequence of the results of Andersen et al. [1]. In Appendix A.3, we proof the claims made in Section 3.3 about the martingale structure of the AV logrank test under the presence of ties. Appendix A.4 shows how to learn the alternative from the data using a Bayesian prior rather than the prequential plug-in approach that was explained in Section 3.2 in the main text. Lastly, in Appendix A.5, we give further details on the simulations used to compute the planned maximum sample sizes for a given targeted power. Under the alternative and optional stopping, the observed sample size is in many cases lower.

A.1 Expected Stopping Time, GROW and Wald's Identity

Here we motivate the 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. Let $\mathbf{P}_0$ represent our null model, and let, as before, the alternative hypothesis be $\mathcal{H}_1 : \theta \leq \theta_1$ for some $\theta_1 < \theta_0$. Suppose we perform a level-α test based on a test martingale $S_{\theta_0,t}^q$ using the stopping rule τ that stops as soon as $S_{\theta_0,t}^q$ exceeds the threshold $1/\alpha$, that is, $\tau^q = \inf_t\{t \ : \ S_{\theta_0,t}^q \geq 1/\alpha\}$. In the main text we elaborated on how $S_{\theta_0,t}^{\theta_1}$ is optimal with respect to the GROW criterion. We now show that the problem of minimizing the worst-case, the expected number of events $\mathbf{E}_\theta[\bar{N}_{\tau^q}]$ over q is approximately equivalent to finding the GROW test martingale. To do so, we make simplifying assumptions that reduce the problem to an i.i.d. experiment. This allows us to employ a standard argument based on an identity of Wald [45], originally due to Breiman [2]. For this we assume that the initial risk sets (i.e., $\bar{y}_0^A$ and $\bar{y}_0^B$) are large enough so that, for all sample sizes we will ever encounter, $\bar{y}_t^A/\bar{y}_t^B \approx \bar{y}_0^A/\bar{y}_0^B$. This allows us to treat the likelihood of the participant(s) $I_{(k)}$ having witnessed the event at time $T^{(k)}$ to be independent of t, that is, as an i.i.d. experiment.

The argument of Breiman [2] relates the expected number of events to the expected value of our stopped AV logrank statistic. Suppose first that we happen to know that the data come from a specific θ in the alternative hypothesis. Then $S_{\theta_0,\tau}^q$ is the product of $\bar{N}_\tau$ factors of ratios $R_{\theta_0,(i)}^q = q_{(i)}(I_{(i)})/p_{\theta_0,(i)}(I_{(i)})$ at the ith event. Wald's iden-

tity applied to its logarithm implies

$$\mathbf{E}_\theta[\bar{N}_\tau] = \frac{\mathbf{E}_\theta[\ln S_{\theta_0,\tau^q}^q]}{\mathbf{E}_\theta[\ln R_{\theta_0,(1)}^q]}. \tag{A.1}$$

For simplicity we will further assume that the number of participants at risk is large enough so that the probability that we run out of data before we can reject is negligible. Because of the choice of the stopping rule τ^q, the right-hand side of the last display can then be further rewritten as

$$\frac{\mathbf{E}_\theta[\ln S_{\theta_0,\tau^q}^q]}{\mathbf{E}_\theta[\ln R_{\theta_0,(1)}^q]} = \frac{\ln(1/\alpha) + \text{VERY SMALL}}{\mathbf{E}_\theta\left[\ln\left(q_{(1)}(I_{(1)})/p_{(1),\theta_0}(I_{(1)})\right)\right]},$$

where VERY SMALL between 0 and $\log|\theta_1/\theta_0|$. The equality follows because we reject as soon as $S_{\theta_0,t}^q \geq 1/\alpha$, so $S_{\theta_0,\tau}^q$ cannot be smaller than $1/\alpha$, and it cannot 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).

With (A.1) at hand, we can relate our choice of q to the expected number of events witnessed before stopping. If, for a fixed θ, we try find the q that minimizes the expected number of events $\mathbf{E}_\theta[\bar{N}_{\tau^q}]$, 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 the numerator $\mathbf{E}_\theta\left[\ln\left(Q_{(1)}/P_{\theta_0,(1)}\right)\right]$ over q. The maximum is achieved by $Q_{(1)} = P_{\theta,(1)}$; the expression in the denominator then becomes the Kulback-Leibler divergence between two Bernoulli distributions. It follows that, under θ, the expected number of outcomes until rejection is minimized by $Q_{(1)} = P_\theta$. Thus, in this case, we use the GROW $S_{\theta_0,t}^\theta$ as test statistic. However, we still need to consider the fact that the real $\mathcal{H}_1$ is composite: as statisticians, we do not know the actual θ; we only know $0 < \theta \leq \theta_1$. A worst-case approach uses the q achieving

$$\max_q \min_{\theta \leq \theta_1} \mathbf{E}_\theta\left[\ln\left(p_{(1)}(I_{(1)})/q_{(1),\theta_0}(I_{(1)})\right)\right]$$

since, repeating the reasoning leading to (A.1), this q should be close to achieving the min-max number of events until rejection, given by

$$\min_q \max_{\theta \leq \theta_1} \mathbf{E}_\theta[\bar{N}_{\tau^q}]$$

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

A.2 Continuous Time and Anytime Validity

In this section, we show the anytime validity of the AV logrank test. This is done via Ville's inequality for which it suffices to show that $S_{\theta_0}^q = (S_{\theta_0,t}^q)_{t\geq 0}$ is a nonnegative (super) martingale. To do so, we use the counting process

formalism. A few definitions are in order. Only in this section, we assume knowledge of counting process theory [see 1, 8]. Denote, for $i = 1, \ldots, m$, $\tilde{N}_t^i = \mathbf{1}\{t \leq T^i\}$ the counting processes associated to each participant, and let y_t^i be the at-risk process. For each participant, the censored process N_t^i, which is observed, is given by $\mathrm{d}N_t^i = y_t^i \mathrm{d}\tilde{N}_t^i$—we use this convention to signify that $N_t^i = \int_0^t y_s^i \mathrm{d}\tilde{N}_s^i$. We define the sigma-algebra $\mathcal{F}_t := \sigma(N_s^j : 0 \leq s \leq t, j = 1, \ldots, n)$, which, as usual, can be interpreted as the information in the study up to time t.

One of the results of the counting process theory is that the processes $\mathrm{d}N_t^i - y_t^i \mathrm{d}\lambda_t^i$ are martingales, where, recall, $y_t^i = \mathbf{1}\{X^i \geq t\}$ is the at-risk process, and λ_t^i is the hazard function associated to T^i. In that case, $y_t^i \mathrm{d}\lambda_t^i$ is called the compensator of N_t^i. The result that the AV logrank test is a martingale hinges specifically on this structure. Thus, any pattern that preserves this martingale structure also preserves the martingale property for the AV logrank test, and consequently its type-I error guarantees. Andersen et al. [1, III.4] show exactly this under general patterns of incomplete observation provided that the mechanisms are independent of the observations. With this in mind, in the following, we only assume that the counting processes N_t^i have compensators A_t^i given by $\mathrm{d}A_t^i = y_t^i \mathrm{d}\lambda_t^i$.

The filtration $\mathcal{F} = (\mathcal{F}_s)_{s \geq 0}$ is right-continuous and we can safely identify predictable processes with left-continuous process. For some θ_0, denote by $\mathbf{P}_0$ the distribution under which, for each $i = 1, \ldots, m$, the hazard function for T^i is $\lambda_t^i = \theta_0^{z^i} \lambda_t^A$, where $g^i = 0$ if $i \in A$ and $g^i = 1$ if $i \in B$. Recall from Section 2, if participant i belongs to Group B, $\lambda_t^i = \theta_0 \lambda_t^B = \theta_0 \lambda_t^A$; otherwise, $\lambda_t^i = \lambda_t^A$. Let $q_t^1, \ldots, q_t^m$ be predictable processes such that $\sum_{i \leq m} q_t^i y_t^i = 1$ a.s. for all t, that is, $\{q_t^i\}_{i \in \mathcal{R}_t}$ at each t is a probability distribution over the participants at risk at time t. Define r_t^i to be each of the ratios $r_t^i = q_t^i / p_{\theta_0, t}^i$. Define the predictable process $S_{\theta_0, t^-}^q = \lim_{s \uparrow t} S_{\theta_0, t^-}^q$. As such, at each t, the change $\mathrm{d}S_{\theta_0, t}^q = S_{\theta_0, t}^q - S_{\theta_0, t^-}^q$ of the AV logrank statistic $S_{\theta_1}^q$ at time t, given in (3.5), can be computed as

$$\mathrm{d}S_{\theta_0, t}^q = \sum_{i \leq m} S_{\theta_0, t^-}^q (r_t^i - 1) \mathrm{d}N_t^i,$$

because no two events happen simultaneously with positive probability. Since S_{θ_0, t^-}^q is predictable, it is enough to prove that the process M_t defined by $\mathrm{d}M_t = \sum_{i \leq m}(1 - r_t^i)\mathrm{d}N_t^i$ is a martingale [see 8, Theorem 1.5.1]. Recall that $\bar{y}_t^A = \sum_{i \in A} y_t^i$ and $\bar{y}_t^B = \sum_{i \in B} y_t^i$. Then both $\bar{y}^A$ and $\bar{y}^B$ are left-continuous processes.

Lemma A.1. *Let $\{q_t^i\}_{i \leq m}$ be a collection of nonnegative left-continuous processes $q^i = (q_t^i)_{t \geq 0}$ such that $\sum_{i \leq m} y_t^i q_t^i = 1$ for all t. Let $\{p_{\theta_0, t}^i\}_{i \leq m}$ be the collection of processes given by*

$$p_{\theta_0, t}^i = \frac{\theta_0^{g^i} y_t^i}{\bar{y}_t^A + \theta_0 \bar{y}_t^B}.$$

The process $M = (M_t)_{t \geq 0}$ given by $\mathrm{d}M_t = \sum_{i \leq m}(1 - r_t^i)\mathrm{d}N_t^i$ is a martingale under $\mathbf{P}_0$ with respect to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geq 0}$.

Proof. It suffices to show that the compensator A_t of M_t, given by $\mathrm{d}A_t = \sum_{i \leq m} \sum_{i \leq m}(r_t^i - 1)y_t^i \lambda_t^i \mathrm{d}t$ is zero. Define $\bar{q}_t^A = \sum_{i \in A} y_t^i q_t^i$ and $\bar{q}_t^B = \sum_{i \in B} y_t^i q_t^i$. Notice that by assumption $\bar{q}_t^A + \bar{q}_t^B = 1.$, and recall that, under the null $\lambda_t^B = \theta_0 \lambda_t^A$. We can compute

$$\sum_{i \leq m}(r_t^i - 1)y_t^i \lambda_t^i = \sum_{i \in A} y_t^i \lambda_t^A(r_t^i - 1) + \sum_{i \in B} y_t^i \lambda_t^B(r_t^i - 1)$$
$$= \lambda_t^A[(\bar{y}_t^A + \theta_0 \bar{y}_t^B)\bar{q}_t^A - \bar{y}_t^A$$
$$+ (\bar{y}_t^A + \theta_0 \bar{y}_t^B)\bar{q}_t^B - \theta_0 \bar{y}_t^B]$$
$$= \lambda_t^A[(\bar{y}_t^A + \theta_0 \bar{y}_t^B)\overbrace{(\bar{q}_t^A + \bar{q}_t^B)}^{=1} - (\bar{y}_t^A + \theta \bar{y}_t^B)]$$
$$= 0,$$

where we used the assumption that $\sum_{i \leq m} y_t^i q_t^i = \bar{y}_t^A q_t^A + \bar{y}_t^B q_t^B = 1$. As the compensator A_t of M_t is zero at each t, we conclude that M_t is a martingale, as was to be shown. $\square$

Our previous discussion and the preceding lemma have the following corollary as a consequence.

Corollary A.2. $S_{\theta_0}^q = (S_{\theta_0, t}^q)_{t \geq 0}$ *is a nonnegative martingale with expected value equal to one.*

Hence, Ville's inequality holds for $S_{\theta_0}^q$, which implies that

$$\mathbf{P}_0\{S_{\theta_0, t}^q \geq 1/\alpha \text{ for some } t \geq 0\} \leq \alpha.$$

This implies the anytime validity of the test $\xi_{\theta_0}^q = (\xi_{\theta_0, t}^q)_{t \geq 0}$ given by the AV logrank test $\xi_{\theta_0, t}^q = \mathbf{1}\{S_{\theta_0, t}^q \geq 1/\alpha\}$.

A.3 Ties

The purpose of this section is twofold. Firstly, we prove Lemma 3.3. Secondly, we show that the conditional likelihood given in Section 3.3 indeed approximates the true conditional partial likelihood ratio under any distribution such that the hazard ratio is θ_1.

Our general strategy in this case is similar to the one undertaken in the continuous-monitoring case: we build a test martingale with respect to a filtration $\mathcal{G}^\star$, and use Ville's inequality to derive anytime-valid type-I error guarantees. Define, for each $k = 1, 2, \ldots$, the sigma-algebra $\mathcal{G}_k$ generated by all observations made in times $t_1, \ldots, t_k$, that is, $\mathcal{G}_k = \sigma(N_{t_l}^i, \tilde{N}_{t_l}^i : i = 1, \ldots, m; l = 1, \ldots, k)$, and the corresponding filtration $\mathcal{G} = (\mathcal{G}_k)_{k=1,2,\ldots}$. Under Cox's proportional hazard model, conditionally on $\mathcal{G}_{k-1}$, our observations $\Delta \bar{N}_k^A$ and $\Delta \bar{N}_k^B$ are binomially distributed with parameters depending on the hazard function (see Lemma A.3 below). By conditioning both on $\mathcal{G}_{k-1}$ and on the total number of events $\Delta \bar{N}_k = \Delta \bar{N}_k^A + \Delta \bar{N}_k^B$, we use the likelihood of having observed $\Delta \bar{N}_k^B$, which follows Fisher's noncentral hypergeometric distribution, as detailed in Corollary A.4. We gather these observations in the following two lemmas.

Lemma A.3. *Conditionally on $\mathcal{G}_{k-1}$, the following hold:*

1. *The number of events $\Delta \bar{N}_k^A$ has a binomial distribution with parameters $\bar{y}_k^A$ and p_k^A where $p_k^A = 1 - \exp(-\int_{t_{k-1}}^{t_k} \lambda_s^A \mathrm{d}s)$.*

2. *The number of events $\Delta \bar{N}_k^B$ has a binomial distribution with parameters $\bar{y}_k^B$ and p_k^B where $p_k^B = 1 - \exp(-\theta \int_{t_{k-1}}^{t_k} \lambda_s^A \mathrm{d}s)$ and θ is the hazard ratio.*

Proof. The result is standard, and it follows from explicitly solving for λ in (2.1) and computing the conditional probability in (2.2) for each group. $\qquad \square$

Next, we use a standard result: given two binomially distributed random variables X and Y, the distribution of X conditionally on $X + Y$ is Fisher's noncentral hypergeometric distribution. We apply this to $\Delta \bar{N}_k^A$ and $\Delta \bar{N}_k^B$ from the previous lemma and spell out the corresponding parameters in the following corollary.

Corollary A.4. *Let $\mathcal{G}_{k-1}^{\star} = \mathcal{G}_{k-1} \vee \sigma(\Delta \bar{N}_k)$, and let p_k^A and p_k^B as in Lemma A.3. Define the odd ratios $\omega_k^A = p_k^A / (1 - p_k^A)$, $\omega_k^B = p_k^B / (1 - p_k^B)$ and the ratio $\omega_k = \omega_k^B / \omega_k^A$. Then, conditionally on $\mathcal{G}_{k-1}^{\star}$, the likelihood of having observed $\Delta \bar{N}_k^B$ events in group B is given by Fisher's noncentral hypergeometric distribution $p_{\mathrm{FNCH}}(\Delta \bar{N}_k^B ; \bar{y}_{k-1}^B, \bar{y}_{k-1}^A, \Delta \bar{N}_k, \omega_k)$, where*

$$p_{\mathrm{FNCH}}(n^B; \bar{y}^B, \bar{y}^A, n, \omega)$$
$$= \frac{\binom{\bar{y}^B}{n^B}\binom{\bar{y}^A}{n-n^B}\omega^{n^B}}{\sum_{\max\{0, n^B - \bar{y}^B\} \leq u \leq \min\{\bar{y}^B, n^B\}} \binom{\bar{y}^B}{u}\binom{\bar{y}^A}{n^B - u}\omega^u}.$$

Naively, one could use a partial likelihood ratio just as in the absence of ties to derive a sequential test. This, however, is not satisfactory, because, in general, the parameter ω_k depends heavily on the unknown baseline hazard function λ^A. Contrary to the general case, when the hazard ratio θ is one, the parameter $\omega_k = 1$, and Fisher's noncentral hypergeometric distribution reduces to the conventional hypergeometric distribution. With this observation at hand, if $\{q_k\}_{k=1,2,\ldots}$ is a sequence of conditional distributions $q_k(\cdot)$ on the possible values of $\Delta \bar{N}_k^B$, we can build a sequential tests for (2.3), with its corresponding type-I error guarantee. We give the details in the following corollary, and subsequently point at a useful choice for q that approximates the real likelihood.

The choice of q for our statistic presented in Section 3.3 follows from an approximation of the parameter ω for small $\Delta t_k = t_k - t_{k-1}$. As noted by [21], if $\int_{t_{k-1}}^{t_k} \lambda_1(s)\mathrm{d}s$ is small, then $p_k^A \approx \lambda_{t_{k-1}} \Delta t_k$ and $p_k^A \approx \theta p_k^A$. With these two approximations, $\omega_k \approx \theta$. This means that the choice $q_k(\Delta \bar{N}_k^B) = p_{\theta_1,k}(\Delta \bar{N}_k^B) := p_{\mathrm{FNCH}}(\Delta \bar{N}_k^B; \bar{y}_k^B, \bar{y}_k^A, \Delta \bar{N}_k, \omega = \theta_1)$ approximates the real conditional likelihood under any alternative for which the true hazard ratio is θ_1. Hence, the se-

quentially computed statistic

$$S_k^{\theta_1} = \prod_{l \leq k} \frac{p_{\theta_1,k}(\Delta \bar{N}_k^B)}{p_{1,k}(\Delta \bar{N}_k^B)}$$

approximates the true partial likelihood ratio of the data observed up to time t_k in the presence of ties, and we recommend its use.

A.4 Bayesian Approach for Numerator

In Section 3.2 we showed how one can deal with alternatives such as $\mathcal{H}_1 : \theta \neq \theta_0$ by learning θ from data, leading to a process $q_{(1)}, q_{(2)}, \ldots$ with $q_{(k)}$ determined by a plug-in estimate of θ. Alternatively, it is also possible to use a Bayes predictive distribution based on a prior $\mathbf{W}$ on θ. If $\mathbf{W}_{(k)} = \mathbf{W} \mid \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}$ is the Bayes posterior on θ based on a prior $\mathbf{W}$ and the data up to time $t < T^{(k)}$, then

$$q_{(k)} = p_{\mathbf{W},(k)} := \int p_{\theta,(k)} \mathrm{d}\mathbf{W}_{(k)}(\theta),$$

where $\mathbf{W}_{(1)} = \mathbf{W}$. Hence, $p_{\mathbf{W},(k)}$ is the Bayesian predictive distribution. The resulting statistic $S_t^{\mathbf{W}}$ is the result of multiplying the conditional probability mass functions $p_{\mathbf{W},(k)}$, and we obtain that

$$S_{\theta_0,t}^{\mathbf{W}} = \prod_{k: T^{(k)} \leq t}^{n} \frac{p_{\mathbf{W},(k)}(I_{(k)})}{p_{\theta_0,(k)}(I_{(k)})} \qquad (\mathrm{A.2})$$

is a Bayes factor between the Bayes marginal distribution based on $\mathbf{W}$ and θ_0. This technique has been employed in sequential analysis; it is known as the method of mixtures [5, 31] and has been compared to the plug-in method discussed in the main text by [30]. We do not know of a prior for which (A.2) or the constituent products have an analytic expression, but it can certainly be implemented using, for example, Gibbs sampling.

As shown in Section 3 and discussed in Section 3.2, the use of any $S_{\theta_0,t}^q$ instead of $S_{\theta,t}^{\theta_1}$ does not compromise on safety: a test based on monitoring $S_{\theta_0}^q$ is anytime-valid, whether q makes reference to plug-in estimators or Bayes predictive distributions, no matter what prior $\mathbf{W}$ was chosen. The type-I error guarantee always holds, also when the prior is "misspecified", putting most of its mass in a region of the parameter space far from the actual θ from which the data were sampled. Thus, our set-up is intimately related to the concept of *luckiness* in the machine learning theory literature [10] rather than to "pure" Bayesian statistics.

A.5 Details of Sample Size Comparison Simulations

In this section we lay out the procedure that we used to estimate the expected and maximum number of events required to achieve a predefined power as shown in Figure 4

and Figure 1 in Section 7. First we describe how we sampled the survival processes under a specific hazard ratio. We then describe how we estimated the maximum and expected sample size required to achieve a predefined power (80% in our case) for any of the test martingales that we considered (that of the exact AV logrank, its Gaussian approximation, and the prequential plugin variant). Finally, we explain how the same quantities for the classical logrank test and the O'Brien-Fleming procedure were obtained.

In order to simulate the order in which the events in a survival processes happens, we used the sequential-multinomial risk-set process from Section 3. As before, we consider the general testing problem with $\theta_0 = 1$ and a minimal clinically relevant effect size $\theta_1 < 1$, and we denote the true data generating parameter by θ, typically, $\theta \leq \theta_1$. Under θ, the odds of the next event at the ith event time happening in Group B are $\theta \bar{y}_{(i)}^B : \bar{y}_{(i)}^A$—the odds change at each time step. Thus, simulating in which group the next event happens only takes a biased coin flip. For the problem of testing (3.6) with θ_0 we fix the tolerate a type-I error to $\alpha = 0.05$ and the type-II error to $\beta = 0.2$. For each test martingale $S_{\theta_0}^q$ of interest we first consider the stopping rule $\tau^q = \inf\{k : S_{\theta_0,(k)}^q \geq 1/\alpha\}$, that is, we stop as soon as $S_{\theta_0,(i)}^q$ crosses the threshold $1/\alpha$. Recall that in the worst case, $\theta = \theta_1$ the expected stopping time τ^q is lowest when we use $S_{\theta_0,(k)}^{\theta_1}$, see Appendix A.1.

To estimate the maximum number of events needed to achieve a predefined power with a given test martingale, we turned our attention to a modified stopping rule $\tilde{\tau}^q$. Under $\tilde{\tau}^q$ we stop at the first of two moments: either when our test martingale $S_{\theta_0,(k)}^q$ crosses the threshold $1/\alpha$ (i.e., at τ) or once we have witnessed a predefined maximum number of events $n_{\max}$. More compactly, this means using the stopping rule $\tilde{\tau}^q$ given by $\tilde{\tau}^q = \min(\tau^q, n_{\max})$. In those cases in which the test based on the stopping rule τ^q achieves a power higher than $1 - \beta$, a maximum number of events $n_{\max}$ smaller than the initial size of the combined risk groups can be selected to achieve approximate power $1 - \beta$ using the rule $\tilde{\tau}^q$.

A quick computation shows that $n_{\max}$ has the following property: it is the smallest number of events n such that stopping after n events has probability smaller than $1 - \beta$ under the alternative hypothesis, that is,

$$\mathbf{P}_\theta\{\tau^q \geq n\} \leq 1 - \beta.$$

More succinctly, $n_{\max}$ is the (approximate) $(1-\beta)$-quantile of the stopping time τ^q, which can be estimated experimentally in a straightforward manner.

To estimate $n_{\max}$ for an initial risk set sizes m_1, m_0, we sampled 10^4 realizations of the survival process (under θ) using the method described at the beginning of this section. This allowed us to obtain the same number of realizations of the stopping time τ^q. We then computed the $(1-\beta)$-quantile of the simulated first passage time distribution of τ^q, and

reported it as an estimate of the number of events $n_{\max}$ in the 'maximum' column in Figure 4.

We assessed the uncertainty in the estimation $n_{\max}$ using the bootstrap. We performed 1000 bootstrap rounds on the sampled empirical distribution of τ^q, and found that the number of realizations that we sampled (10^4) was high enough so that plotting the uncertainty estimates was not meaningful relative to the scale of our plots. For this reason we omitted the error bars in Figure 4 and Figure 1.

In the "mean" column of Figure 4 and Figure 1 we plot an estimate of the expected number of events $\tilde{\tau}^q = \min(\tau^q, n_{\max})$. For this, we used the empirical mean of the stopping times that were smaller than $n_{\max}$ on the sample that we obtained by simulation, with 20% of the stopping times being $n_{\max}$ itself. In the "conditional mean" column, we plot an estimate of $\tilde{\tau}^q \mid \tilde{\tau}^q < n_{\max}$, i.e., the stopping time given that we stop early (and hence reject the null).

For comparison, we also show the number of events that one would need under the Gaussian non-sequential approximation of Schoenfeld [32], and under the continuous monitoring version of the O'Brien-Fleming procedure. In order to judge Schoenfeld's approximation, we report the number of events required to achieve 80% power. This is equivalent to treating the logrank statistic as if it were normally distributed, and rejecting the null hypothesis using a z-test for a fixed number of events. The power analysis of this procedure is classic, and the number of events required is $n_{\max}^S = 4(z_\alpha + z_\beta)^2 / \log^2 \theta_1$, where z_α, and z_β are the α, and β-quantiles of the standard normal distribution. In the case of the continuous monitoring version of O'Brien-Fleming's procedure, we estimated the number of events $n_{\max}^{OF}$ needed to achieve 80% as follows. For each experimental setting (m^A, m^B, θ), we generated 10^4 realizations of the survival process under θ and computed the corresponding trajectories of the logrank statistic. For each possible value n of $n_{\max}^{OF}$, we computed the fraction of trajectories for which the O'Brien-Fleming procedure correctly stopped when used with the maximum number of events set to n. We report as an estimate of the true $n_{\max}^{OF}$ the first value of n for which this fraction is higher than 80%, our predefined power.

APPENDIX B. COVARIATES: THE FULL COX PROPORTIONAL HAZARDS *E*-VARIABLE

In the concluding Section 8.1, we indicated how to extend our approach to the full Cox' proportional hazards model with covariates, using the *Universal Inference* ('running MLE') approach. Here we show that we may also proceed using the *Reverse Information Projection (RIPr)* approach pioneered by [9]. Our starting point are the partial likelihoods (8.1).

Since the RIPr approach inevitably requires the use of prior distributions on the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ appearing in $\mathcal{H}_0$, it is more convenient to also treat composite alternatives $\mathcal{H}_1$ in a Bayesian manner as described in Appendix A.4

rather than using the prequential plug-in approach of Section 3.2, so for simplicity, below we will only describe this approach. To be sure though, we may use the plug-in approach here as well, if we are so inclined.

B.1 *E*-Variables and Martingales

Let $\mathbf{W}$ be a prior distribution on $\beta \in \mathbb{R}^d$ for some $d > 0$. ($\mathbf{W}$ may be degenerate, i.e., put mass one on a specific parameter vector β_1, and this will allow us to deal with the analogue of the one-sided tests of Section 2 rather than just 'learning the alternative distribution' tests of Section 3.2). For each such $\mathbf{W}$, we let $q_{\mathbf{W},\theta,(k)}$ be the probability distribution on $\mathcal{R}_{(k)}$ defined by

$$q_{\mathbf{W},\theta,(k)}(\ i\ ) := \int p_{\beta,\theta,(k)}(\ i\ )\mathrm{d}\mathbf{W}(\beta),$$

with $p_{\beta,\theta,(k)}$ as in (8.1). Consider a measure ρ on $\mathbb{R}^d$ (e.g., Lebesgue or some counting measure) and we let $\mathcal{W}$ be the set of all distributions on $\mathbb{R}^d$ 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 $\tilde{q}_{\leftarrow\mathbf{W},\theta_0}$ to be the *reverse information projection* [18] (RIPr) of $q_{\mathbf{W},\theta,(k)}$ on $\{q_{\mathbf{W},\theta_0,(k)} : \mathbf{W} \in \mathcal{W}^\circ\}$, defined as the probability distribution on $\mathcal{R}_{(k)}$ such that

$$\begin{aligned}&\mathrm{KL}(q_{\mathbf{W},\theta_1,(k)}\|\tilde{q}_{\leftarrow\mathbf{W},\theta_0,(k)})\\&\quad = \inf_{\mathbf{W}^\circ \in \mathcal{W}^\circ} \mathrm{KL}(q_{\mathbf{W},\theta_1,(k)}\|q_{\mathbf{W}^\circ,\theta_0,(k)}).\end{aligned}$$

We know from Li [18] and Grünwald et al. [9] that $\tilde{q}_{\leftarrow\mathbf{W},\theta_0,(k)}$ exists for each k. Grünwald et al. [9] show, in the context of *E*-variables for 2×2 contingency tables, that the infimum in the previous display is in fact achieved by some distribution $\mathbf{W}^\star$ with finite support on $\mathbb{R}^d$ if the random variables $y^1_{(k)}, \ldots, y^m_{(k)}$ constituting our random process have a finite range. For given hazard ratios $\theta_0, \theta_1 > 0$, let

$$R^{\theta_1}_{\mathbf{W},\theta_0,(k)} = \frac{q_{\mathbf{W},\theta_1,(k)}(I_{(k)})}{q_{\leftarrow\mathbf{W},\theta_0,(k)}(I_{(k)})} \tag{B.1}$$

be our analogue of (3.3).

Theorem B.1 (Corollary of Theorem 1 from Grünwald et al. [9]). *For every prior $\mathbf{W}$ on $\mathbb{R}^d$, for all $\beta \in \mathbb{R}^d$,*

$$\begin{aligned}&\mathbf{E}_{\beta,\theta_0}[R^{\theta_1}_{\mathbf{W},\theta_0,(k)} \mid \mathbf{z}^i_{(l)}, y^i_{(l)};\ i = 1,\ldots,m;\ l = 1,\ldots,k]\\&\quad = \sum_{i \in \mathcal{R}_{(k)}} q_{\beta,\theta_0,(k)}(i)\frac{q_{\mathbf{W},\theta_1,(k)}(i)}{q_{\leftarrow\mathbf{W},\theta_0,(k)}(i)} \leq 1\end{aligned}$$

so that $R^{\theta_1}_{\mathbf{W},\theta_0,(k)}$ is an E-variable conditionally on $\mathbf{z}^i_{(l)}$, $y^i_{(l)}$ with $i = 1,\ldots,m; l = 1,\ldots,k$.

Note that the result does not require the prior $\mathbf{W}$ to be well specified in any way: under any (β, θ_0) in the null distribution, even if β is completely disconnected to $\mathbf{W}$, $R^{\theta_1}_{\mathbf{W},\theta_0,(k)}$ is an *E*-variable conditional on past data.

In particular, since the result holds for arbitrary priors, it holds, at the kth event time, for the Bayesian posterior $\mathbf{W}_{k+1} = \mathbf{W}_1 \mid \mathbf{z}^i_{(l)}, y^i_{(l)};\ i = 1,\ldots,m;\ l = 1,\ldots,k$, based on arbitrary prior $\mathbf{W}_1$ with density w_1, i.e., the density of $\mathbf{W}_{k+1}$ is given by

$$w_{k+1}(\beta) \propto \prod_{l \leq k} q_{\beta,\theta,(l)}(I_{(l)})w_1(\beta).$$

In parallel to the discussion in Section 3.1, we can therefore, for each prior $\mathbf{W}_1$, construct a test martingale $S_k := \prod_{l \leq k} R^{\theta_1}_{\mathbf{W}_l,\theta_0,(l)}$ that "learns" β from the data, analogously to (A.2), and computes a new RIPr at each event time k.

B.2 Finding the RIPr

While it is not clear how to calculate the RIPr $q_{\leftarrow\mathbf{W},\theta_0,(k)}$ in general, it can be well approximated with the efficient algorithm design by Li [18] and Li and Barron [17]. Their algorithm is computationally feasible as long as we restrict $\mathbf{W}^\circ_\delta$ to be the set of all priors $\mathbf{W}$ for which $\min_{i \in \mathcal{R}_{(k)}} q_{\mathbf{W},\theta_0,(k)}(i) \geq \delta$, for some $\delta > 0$. In that case, when run for M steps, the algorithm achieves an approximation error of $O(\ln(1/\delta)/M)$, where each step is linear in the dimension d. Since the approximation error is logarithmic in $1/\delta$, we can take a very small value of δ, which makes the requirement less restrictive. Exploring whether the Li-Barron algorithm really allows us to compute the RIPr for the Cox model, and hence $R^{\theta_1}_{\mathbf{W}_k,\theta_0,(k)}$ in practice, is a major goal for future work.

B.3 Ties

Without covariates, our E-variables allow for ties correspond to a likelihood ratio of Fisher's noncentral hypergeometric distributions (see Section 3.3), the situation is not so simple in the presence of covariates. Although deriving the appropriate extension of the noncentral hypergeometric partial likelihood is possible, one ends up with a hard-to-calculate formula [23]. Various approximations have been proposed in the literature [3, 7]. 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.

APPENDIX C. GAUSSIAN AV LOGRANK TEST

In this section we heuristically derive the Gaussian AV logrank test of Section 4, and investigate the validity of the Gaussian approximation. In Appendix C.1, we show by simulation that this approximation is only valid when the allocation of participants to each group under investigation

is balanced, that is, when $m^A = m^B$. In Appendix C.2 we investigate numerically the sample size needed to reject the null hypothesis under both the exact AV logrank test and its Gaussian approximation.

We start with the derivation of (4.3). For this we use (local) asymptotic normality of the Z-score (4.1). Under the null distribution, Z_k from (4.1) has an asymptotic standard Gaussian distribution. Under any alternative distribution under which the hazard ratio is θ, Schoenfeld [32] showed that, in the absence of ties, the Z-statistic also asymptotically follows a Gaussian distribution with unit variance, but this time with mean $\mu_1^\star$ given by

$$\mu_1^\star = \frac{\sum_{i \leq k} E_i^B (1 - E_i^B)}{\sqrt{\sum_{i \leq k} E_i^B (1 - E_i^B)}} \log(\theta).$$

In the above, 'asymptotically' means in the limit $m^A \to \infty$, $m^B = \gamma m^A$ for some fixed γ, $k \to \infty$. $\log \theta = O((m^A + m^B)^{-1/2})$. Note that $\mu_1^\star$ depends on more than the summary statistic Z_k. In the case that the number of observed events is much smaller than the initial risk set sizes (i.e. additionally we require $k/m^A \to 0$), the mean $\mu_1^\star$ under the alternative can be further approximated by

$$\mu_1^\star \approx \sqrt{\bar{N}_k} \mu_1 = \sqrt{\bar{N}_k} \sqrt{\frac{m^B m^A}{(m^B + m^A)^2}} \log(\theta), \quad \text{(C.1)}$$

where $\bar{N}_k$ is the total number of observations up until time t_k, and the resulting approximation only depends on summary statistics. It is exactly this value μ_1 that we use in the Gaussian AV logrank test. Inspecting the proof of the asymptotic result of Schoenfeld, we find that it relies on two conditions: (1) that the hazard ratio θ_1 under the alternative is close enough to one so that a first-order Taylor approximation around $\theta_0 = 1$ is adequate; (2) that the expected number of events E_k^B stays approximately constant over time, that is, close to the initial allocation proportion $E_1^B = m^B/(m^B + m^A)$. This indicates that the asymptotic approximation is reasonable for values of θ_1 close to 1 and the initial risk sets are both large in comparison to the number of events witnessed. Notice that in this regime of large risk sets the multiplicity correction in V_k is also negligible.

This raises the question whether a sequential Gaussian approximation is sensible for the logrank statistic—a priori it is not at all clear whether Schoenfeld's asymptotic fixed-sample result still provides a reasonable approximation for the partial likelihood ratio under optional stopping. We now investigate this question empirically (as remarked by a referee, it may be that the techniques of [50] on time-uniform central limit theory could be extended to investigate this more rigorously, but the details being far from evident, we have refrained from doing so). Define the logrank statistic per observation time

$$Z_i = \frac{O_i^B - E_i^B}{\sqrt{V_i^B}}.$$

We investigate whether the exact AV logrank statistic behaves similarly to the Gaussian likelihood ratio

$$S_k'^G = \prod_{i \leq k} \frac{\phi_{\mu_1 \sqrt{O_i}}(Z_i)}{\phi_{\mu_0}(Z_i)}$$

$$= \exp\left(-\frac{1}{2} \sum_{i \leq k} \left\{ O_i \mu_1^2 - 2\mu_1 \sqrt{O_i} Z_i \right\} \right)$$

for $\theta_0 = 1$ we have $\mu_0 = 0$, $\mu_1 = \log(\theta) \sqrt{m^B m^A / (m^A + m^B)^2}$, and ϕ_μ is the Gaussian density with unit variance and mean μ. Note that the statistic still depends on elements of the full data set; more approximations are needed. Write the Gaussian densities, and use that in the limit of large risk sets $p_i^B \approx m^B/(m^A + m^B)$ and that consequently $V_i \approx \sqrt{O_i \frac{m^A m^B}{(m^A + m^B)^2}}$. These approximations are valid under Schoenfeld's second assumption. With these approximations at hand, the Z-statistic is approximated by

$$Z_k \approx \frac{\sum_{i \leq k} \left\{ O_i^B - E_i^B \right\}}{\sqrt{O_i \frac{m^A m^B}{(m^A + m^B)^2}}}$$

and consequently

$$S_k'^G \approx S_k^G = \exp\left(-\frac{1}{2} \bar{N}_k \mu_1^2 + \sqrt{\bar{N}_k} \mu_1 Z_k \right),$$

where S_k^G is as in (4.3). In Figure 5 we show, in case of balanced allocation, that the Gaussian approximation S_k^G at a single event time is very similar to the exact $S_{\theta_0,(k)}^{\theta_1}$ for alternative hazard ratios θ_1 between 0.5 and 2.

C.1 Safety Only for Balanced Allocation

In order to assess whether the Gaussian AV logrank test is indeed AV, that is, whether the type-I error guarantees holds, we inspect whether the expected value of each of its multiplicative increments is bellow 1. In relation to our discussion in Section 3.1, this would imply that all multiplicative increments are conditional E-variables and that the resulting test is, at least approximately, a test martingale. Figure 6 shows the expectation of these increments as a function of the hazard ratio for several initial allocation ratios. In case of balanced 1:1 allocation S_k^G is an E-variable, since its expectation is 1 or smaller. However, in case of unbalanced 2:1 or 3:1 allocation and designs with hazard ratio $\theta_1 < 1$, S_k^G is not an E-variable. Of course, even if the initial allocation is balanced, it can become unbalanced. Figure 6 shows that in case of designs outside the range $0.5 \leq \theta_1 \leq 2$ the deviations from expectation 1 can be problematic. Hence we do not recommend to use the Gaussian approximation on the logrank statistic for unbalanced designs and designs for $\theta_1 < 0.5$ or $\theta_1 > 2$. For balanced designs with $0.5 \leq \theta_1 \leq 2$, we found that in practice they are safe to use, the reason
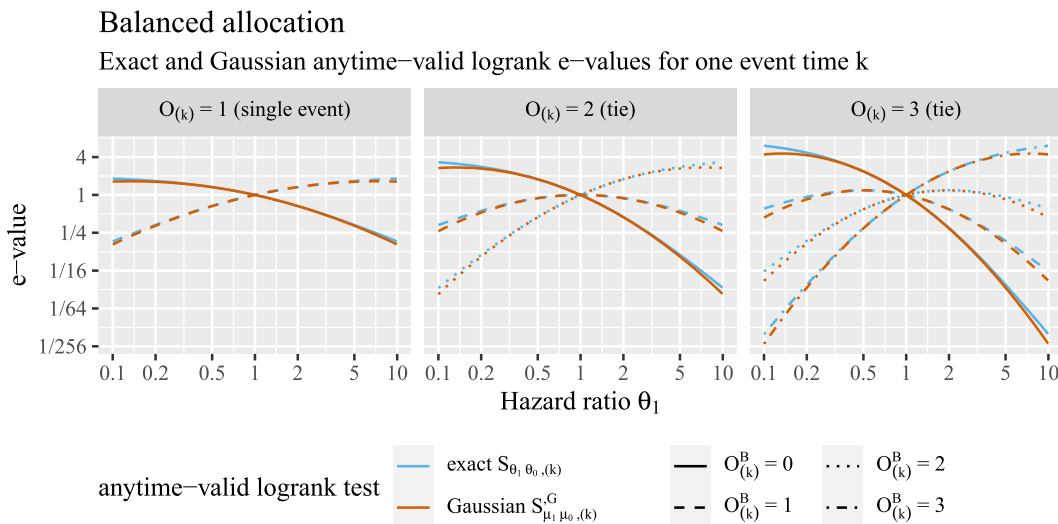
Balanced allocation

Exact and Gaussian anytime−valid logrank e−values for one event time k

Figure 5: For balanced allocation $(m^A = m^B)$ $S_1'^G$ is very similar to $S_{(1)}$ when $0.5 \leq \theta_1 \leq 2$. Here $\theta_0 = 1$, $\mu_0 = 0$, and $\mu_1 = \mu_1(\theta_1)$ as in in (C.1). Note that both axis are logarithmic.
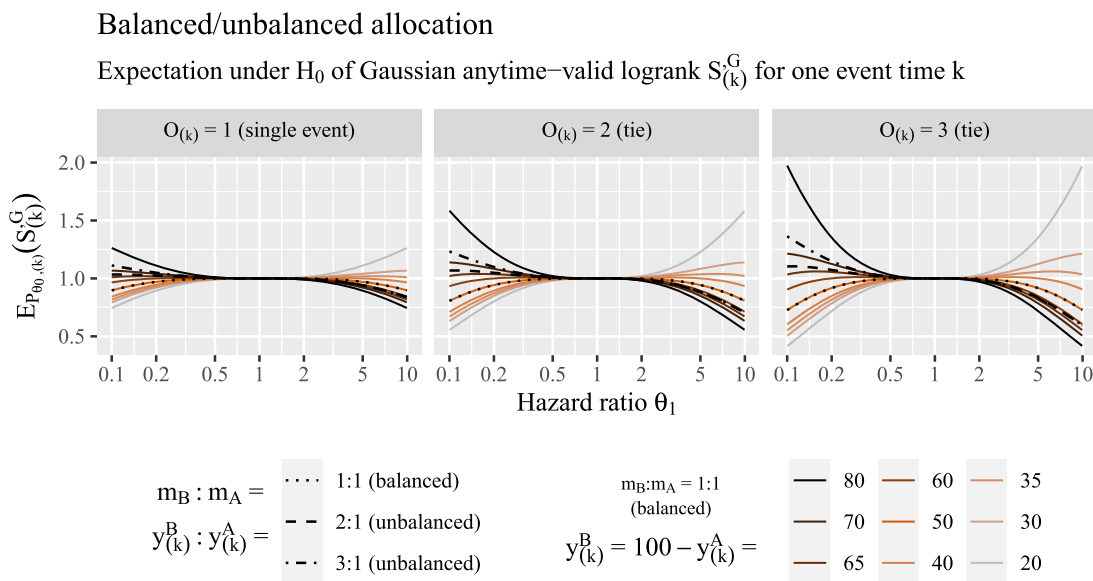
Balanced/unbalanced allocation

Expectation under H_0 of Gaussian anytime−valid logrank $S_{(k)}^G$ for one event time k

Figure 6: Expected value of the increments of the Gaussian AV logrank statistic as a function of the hazard ratio θ_1. For balanced allocation R_i^G is an E-variable, but it is not for unbalanced allocation. The risk set can also start out balanced but become unbalanced; this is unlikely under the null hypothesis (see Appendix C.1). Note that the x-axis is logarithmic.

being that scenarios in which the allocation becomes highly unbalanced after some time (e.g. $y_i^B = 80$, $y_i^A = 20$) are extremely unlikely to occur under the null.

C.2 Sample Size

In this section we compare the stopping time distribution $\tau^G := \inf\{k : \xi_k^G = 1\}$ of the Gaussian approximation to that of $\tau = \inf\{k : \xi_k = 1\}$. We use tests with tolerable type I error $\alpha = 0.05$, thus, the threshold $1/\alpha = 20$ for both tests. In the previous section we showed that the Gaussian

approximation to the AV logrank statistic is valid when the initial allocation is 1:1 and for values $0.5 \leq \theta_1 \leq 2$, where θ_1 is the hazard ratio under the alternative. In these scenarios, we simulate a survival process from a distribution according to which the true data generating hazard ratio is $\theta = \theta_1$ and sampled realizations τ^G and τ for the same data set. The results of the simulation are shown in Figure 7, where we plot the realizations of τ^G against those of τ. We see that in most cases both tests reject at the same time $\tau^G = \tau$, and that the approximation becomes better as θ_1 moves closer

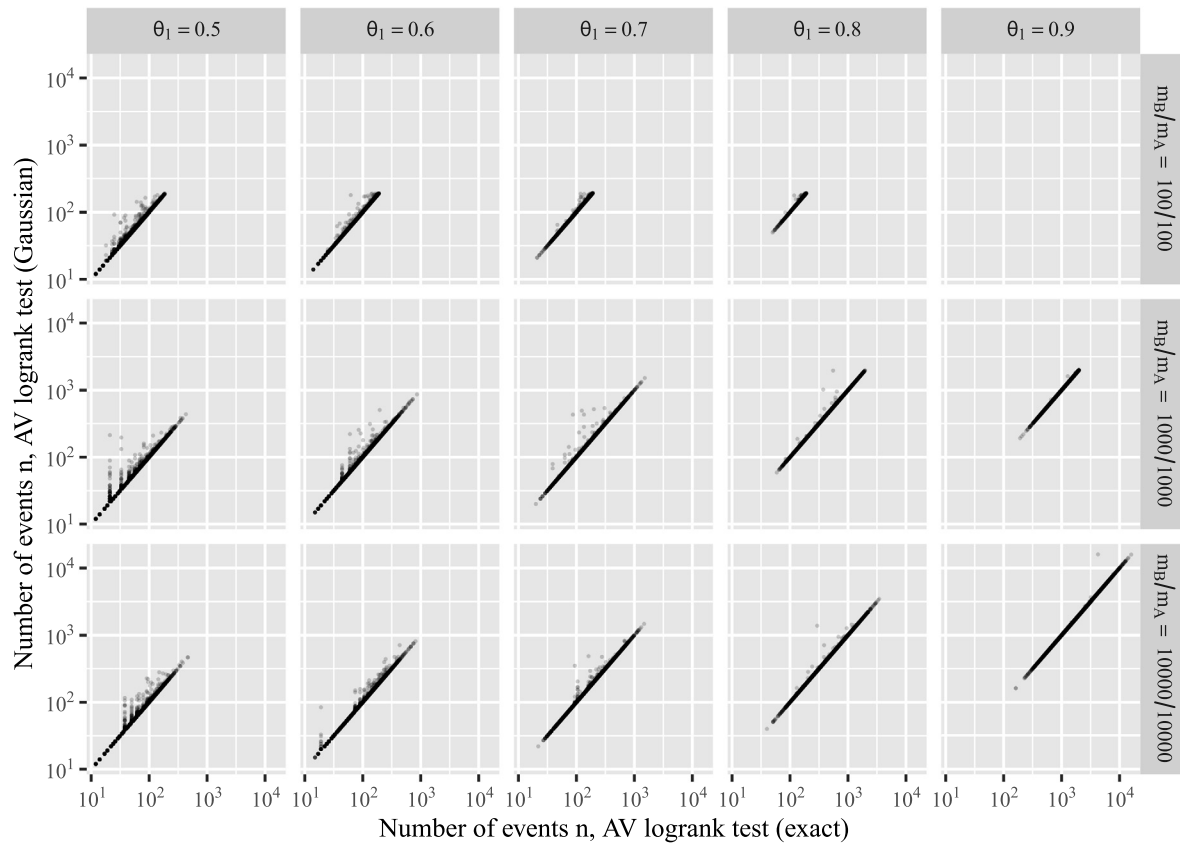## Stopping times for the Gaussian and the exact AV logrank test

Figure 7: Stopping times for the Gaussian and exact AV logrank tests under continuous monitoring (no ties) with threshold $1/\alpha = 20$. The stopping times under the Gaussian approximation often coincide with the exact ones, and are often more conservative (see Appendix C.2). Note that both axes are logarithmic.

to $\theta_0 = 1$ (Schoenfeld's assumption 1). When both tests do not reject at the same time, the Gaussian approximation errs on the conservative side. The deviations from the constant large and balanced risk set do not seem to occur often for this range of hazard ratios. After all, the risk set needs to be large to observe the number of events to detect hazard ratios in the range $0.5 \leq \theta_1 \leq 2$.

REFERENCES

[1] ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. *Statistical Models Based on Counting Processes.* Springer (1993). ISBN . Google-Books-ID: fh3SBwAAQBAJ.

[2] BREIMAN, L. Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, vol. **4.1**, pp. 65–79. University of California Press (January 1961). MR0135630. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fourth-Berkeley-Symposium-on/chapter/Optimal-Gambling-Systems-for-Favorable-Games/bsmsp/1200512159.

[3] COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2) 187–220 (1972). ISSN 0035-9246. Publisher: [Royal Statistical Society, Wiley]. https://www.jstor.org/stable/2985181. MR0341758

[4] COX, D. R. Partial likelihood. *Biometrika* **62**(2) 269–276 (1975). ISSN 0006-3444. Publisher: [Oxford University Press, Biometrika Trust]. https://doi.org/10.1093/biomet/62.2.269. https://doi.org/10.2307/2335362. MR0400509

[5] DARLING, D. A. and ROBBINS, H. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences* **58**(1) 66–68 (July 1967). Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.58.1.66. MR0215406

[6] DAWID, A. P. Present position and potential developments: some personal views: statistical theory: the prequential approach. *Journal of the Royal Statistical Society. Series A (General)* **147**(2) 278–292 (1984). ISSN 0035-9238. Publisher: [Royal Statistical So-

ciety, Wiley]. https://doi.org/10.2307/2981683. MR0763811

[7] EFRON, B. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**(359) 557–565 (September 1977). ISSN 0162-1459. https://doi.org/10.1080/01621459.1977.10480613. MR0451514

[8] FLEMING, T. R. and HARRINGTON, D. P. *Counting Processes and Survival Analysis.* John Wiley & Sons (September 2011). ISBN . MR1100924

[9] GRÜNWALD, P. D., DE HEIDE, R. and KOOLEN, W. Safe testing. *Journal of the Royal Statistical Society, Series B* (2024). With Discussion. arXiv:1906.07801.

[10] GRÜNWALD, P. D. and MEHTA, N. A. A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pp. 433–465. PMLR (March 2019). ISSN 2640-3498. https://proceedings.mlr.press/v98/grunwald19a.html. MR3932854

[11] GRÜNWALD, P. D. and ROOS, T. Minimum description length revisited. *International Journal of Mathematics for Industry* **11**(01) 1930001 (December 2019). ISSN 2661-3352. Publisher: World Scientific Publishing Co. https://doi.org/10.1142/s2661335219300018. MR4090761

[12] HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys* **17** 257–317 (2020). https://doi.org/10.1214/18-PS321. MR4100718

[13] HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* **49**(2) 1055–1080 (2021). https://doi.org/10.1214/20-aos1991. MR4255119

[14] KIM, K. and DEMETS, D. L. Confidence intervals following group sequential tests in clinical trials. *Biometrics* **43**(4) 857–864 (1987). ISSN 0006-341X. Publisher: [Wiley, International Biometric Society]. https://doi.org/10.2307/2531539. MR0920470

[15] KLEIN, J. P. and MOESCHBERGER, M. L. *Survival Analysis: Techniques for Censored and Truncated Data. Statistics for Biology and Health.* Springer, New York, NY (2003). ISBN , 978-0-387-21645-4. https://doi.org/10.1007/b97377

[16] TZE LEUNG LAI. On confidence sequences. *The Annals of Statistics* **4**(2) 265–280 (March 1976). ISSN 0090-5364, 2168-8966. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176343406. MR0395103.

[17] LI, J. and BARRON, A. Mixture density estimation. In *Advances in Neural Information Processing Systems*, vol. **12**. MIT Press (1999). https://papers.nips.cc/paper/1999/hash/a0f3601dc682036423013a5d965db9aa-Abstract.html.

[18] LI, Q. (J.) *Estimation of Mixture Models.* PhD Thesis, Yale University, New Haven, CT, USA, 1999. MR2699116

[19] LINDON, M. and MALEK, A. Anytime-valid inference for multinomial count data. In (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh, eds.) *Advances in Neural Information Processing Systems*, vol. **35**, pp. 2817–2831. Curran Associates, Inc. (2022). https://proceedings.neurips.cc/paper_files/paper/2022/file/12f3bd5d2b7d93eadc1bf508a0872dc2-Paper-Conference.pdf.

[20] MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**(3) 163–170 (March 1966). ISSN 0069-0112.

[21] MEHROTRA, D. V. and ROTH, A. J. Relative risk estimation and inference using a generalized logrank statistic. *Statistics in Medicine* **20**(14) 2099–2113 (2001). ISSN 1097-0258. https://doi.org/10.1002/sim.854.

[22] O'BRIEN, P. C. and FLEMING, T. R. A multiple testing procedure for clinical trials. *Biometrics* **35**(3) 549–556 (1979). ISSN 0006-341X. Publisher: [Wiley, International Biometric Society]. https://doi.org/10.2307/2530245

[23] PETO, R. Discussion of: Regression models and life tables, by DR Cox. *Journal of the Royal Statistical Society, Series B* **26** 205–207 (1972). MR0341758

[24] PETO, R. and PETO, J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)* **135**(2) 185–207 (1972). ISSN 0035-9238. Publisher: [Royal Statistical Society, Wiley]. https://doi.org/10.2307/2344317

[25] POCOCK, S. J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**(2) 191–199 (August 1977). ISSN 0006-3444. https://doi.org/10.1093/biomet/64.2.191

[26] POCOCK, S. J. Current controversies in data monitoring for clinical trials. *Clinical Trials* **3**(6) 513–521 (December 2006). ISSN 1740-7745. Publisher: SAGE Publications https://doi.org/10.1177/1740774506073467.

[27] PROSCHAN, M. A., LAN, K. K. G. and TURK WITTES, J. *Statistical Monitoring of Clinical Trials: A Unified Approach.* Springer (December 2006). ISBN . Google-Books-ID: BCu8c8NwXxcC.

[28] RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv:2009.03167 [math, stat] (September 2020).

[29] RAMDAS, A., GRÜNWALD, P., VOVK, V. and SHAFER, G. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.* **38**(4) 576–601 (2023). ISSN 0883-4237. https://doi.org/10.1214/23-sts894. MR4665027

[30] ROBBINS, H. and SIEGMUND, D. The expected sample size of some tests of power one. *The Annals of Statistics* **2**(3) 415–436 (May 1974). ISSN 0090-5364. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176342704. MR0448750

[31] ROBBINS, H. and SIEGMUND, D. Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics* **41**(5) 1410–1429 (October 1970). ISSN 0003-4851, 2168-8990. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aoms/1177696787. MR0277059

[32] SCHOENFELD, D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**(1) 316–319 (1981). ISSN 0006-3444. Publisher: [Oxford University Press, Biometrika Trust]. https://doi.org/10.1093/biomet/68.1.316. MR0614969

[33] TER SCHURE, J. A. (JUDITH), LY, A., BELIN, L., BENN, C. S., BONTEN, M. J. M., CIRILLO, J. D., DAMEN, J. A. A., FRONTEIRA, I., HENDRIKS, K. D., JUNQUEIRA-KIPNIS, A. P., KIPNIS, A., LAUNAY, O., MENDEZ-REYES, J. E., NETEA, M. G., NIELSEN, S., UPTON, C. M., VAN DEN HOOGEN, G., WEEHUIZEN, J. M., GRÜNWALD, P. D. and VAN WERKHOVEN, C. H. (HENRI). Bacillus Calmette-Guérin vaccine to reduce COVID-19 infections and hospitalisations in healthcare workers: a living systematic review and prospective ALL-IN meta-analysis of individual participant data from randomised controlled trials. medRxiv (2022). https://doi.org/10.1101/2022.12.15.22283474

[34] TER SCHURE, J. and GRÜNWALD, P. Accumulation Bias in meta-analysis: the need to consider time in error control [version 1; peer review: 2 approved]. F1000Research, 8:962, June 2019. ISSN 2046-1402. https://f1000research.com/articles/8-962/v1.

[35] TER SCHURE, J. and GRÜNWALD, P. ALL-IN meta-analysis: breathing life into living systematic reviews [version 1; peer review: 1 approved, 2 approved with reservations]. F1000Research, page 11:549, 2022. https://f1000research.com/articles/11-549.

[36] SELLKE, T. and SIEGMUND, D. Sequential analysis of the proportional hazards model. *Biometrika* **70**(2) 315–326 (1983). Publisher: Oxford University Press. https://doi.org/10.1093/biomet/70.2.315. MR0712020

[37] SHAFER, G. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A* **184**(2) 407–431 (2021). Publisher: Royal Statistical Society. https://doi.org/10.1111/rssa.12647. MR4255905

[38] SLUD, E. V. Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics* **12**(2) 551–571 (June 1984). ISSN 0090-5364, 2168-8966. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176346505. MR0740911

[39] SLUD, E. V. Partial likelihood for continuous-time stochastic processes. *Scandinavian Journal of Statistics* **19**(2) 97–109 (1992).

ISSN 0303-6898. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley]. https://www.jstor.org/stable/4616231. MR1173593

[40] Tse, T. and Davison, A. C. A note on universal inference. *Stat* **11**(1) e501 (2022). https://doi.org/10.1002/sta4.501. MR4529724

[41] Tsiatis, A. A. A large sample study of Cox's regression model. *The Annals of Statistics* **9**(1) 93–108 (January 1981). ISSN 0090-5364, 2168-8966. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176345335

[42] Tsiatis, A. A. *Group Sequential Methods for Survival Analysis with Staggered Entry*. Institute of Mathematical Statistics (1982). ISBN . https://doi.org/10.1214/lnms/1215464854. MR0734207

[43] Turner, R., Ly, A., Perez-Ortiz, M. F., ter Schure, J. and Grunwald, P. D. safestats: Safe Anytime-Valid Inference, November 2022. https://CRAN.R-project.org/package=safestats.

[44] Ville, J. *Etude critique de la notion de collectif*. Gauthier-Villars (1939). MR3533075

[45] Wald, A. *Sequential Analysis*. Wiley, New York (1947). MR0020764

[46] Wang, R. and Ramdas, A. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84**(3) 822–852 (2022). MR4460577

[47] Wasserman, L., Ramdas, A. and Balakrishnan, S. Universal inference. *Proceedings of the National Academy of Sciences* **117**(29) 16880–16890 (July 2020). ISSN 0027-8424, 1091-6490. Publisher: National Academy of Sciences Section: Physical Sciences. https://doi.org/10.1073/pnas.1922664117

[48] Waudby-Smith, I. and Ramdas, A. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, vol. **33**, pp. 20204–20214 (2020).

[49] Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2023). ISSN 1369-7412. https://doi.org/10.1093/jrsssb/qkad009. MR4716192

[50] Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H. and Ramdas, A. Time-uniform central limit theory and asymptotic confidence sequences. arXiv:2103.06476 (2021).

[51] Wu, J. and Xiong, X. Group sequential survival trial design and monitoring using the log-rank test. *Statistics in Biopharmaceutical Research* **9**(1) 35–43 (January 2017). https://doi.org/10.1080/19466315.2016.1189355

Judith ter Schure. Amsterdam University Medical Centre, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. E-mail address: `j.a.terschure@amsterdamumc.nl`

Muriel F. Pérez-Ortiz. Statistics, Probability and Operations Research cluster, Eindhoven University of Technology, Den Dolech 2, 5612 AZ Eindhoven, The Netherlands. E-mail address: `m.f.perez.ortiz@tue.nl`

Alexander Ly. Machine Learning Group, CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands. E-mail address: `a.ly@cwi.nl`

Peter D. Grünwald. Machine Learning Group, CWI (also affiliated with Mathematical Institute, Leiden Unversity), Science Park 123, 1098 XG Amsterdam, The Netherlands. E-mail address: `pdg@cwi.nl`