ALL-IN meta-analysis

Schure, J.A. ter

2 | The Safe logrank test

Abstract

We introduce the safe 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. All these properties are a virtue of the recently developed martingale tests based on E-variables, of which the safe logrank test is an instance. We demonstrate the validity of the underlying nonnegative martingale in a semi-parametric setting of proportional hazards and show how to extend it to ties, Cox' regression and confidence sequences. Using a Gaussian approximation on the logrank statistic, we show that the safe 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.

Introduction

Traditional hypothesis tests and confidence intervals lose their type-I error and coverage guarantees, thus, validity and interpretability, under *optional stopping* and *continuation*. Roughly, optional stopping refers to stopping earlier than originally planned, for example, when results look good enough; optional continuation refers to adding additional data at the end of a trial, or even starting a new trial and combining results, for example, when results of the trial are promising but not fully conclusive.

Recently, a new theory of testing and estimation has emerged for which optional stopping and continuation pose no problem at all (Shafer et al., 2011; Howard et al., 2021; Ramdas et al., 2020; Vovk and Wang, 2021; Shafer, 2021; Grünwald et al., 2019; Turner et al., 2021). The main ingredients are (1) the E-variable that has as outcome an e-value, a direct alternative to the classical p-value, and (2) the test martingale, a product of conditional E-variables. Both are used to create so-called *safe* tests that provide type-I error control under optional stopping and optional continuation, and *anytime-valid* confidence intervals that remain valid irrespective of the stopping time employed. Pace and Salvan

(2019) argue that even without optional stopping, anytime-valid confidence intervals may be preferable over standard ones. Here we provide a concrete instance of this theory: we develop E-variables and martingales for a safe version of the classical logrank test of survival analysis (Mantel, 1966; Peto and Peto, 1972): safe under both optional stopping and continuation. The E-variables, martingales and the corresponding tests are implemented in the SafeStats R package (Turner et al., 2022).

The logrank test is often used in randomized clinical trials to test a difference between two groups in survival time or other time to an event. Its test statistic appears as the score test corresponding to the Cox (1972) proportional hazards model – when the only covariate is the treatment/control indicator – and is also a key tool in statistical monitoring of trials by means of group sequential/α-spending approaches. These sequential methods allow several interim looks at the data to stop for efficacy or futility. Like ours, they are connected to early work by H. Robbins and his students (Darling and Robbins, 1967; Lai, 1976), but the details are very different and in some cases, our approach is more straightforward. In case of unbalanced allocation, for example, α-spending approaches do not provide strong type-I error guarantees (Wu and Xiong, 2017) due to the approximations involved. The basic version of the safe logrank test, however, is exact, without any approximations, so that unbalanced allocation is no problem at all. This ties to the important advantage of using E-variables instead of α-spending: it is more flexible, and as a consequence, easier to use.

Group sequential approaches require prespecified interim looks, in terms of the information fraction of the trial; α-spending is less rigid but still needs a maximum sample size to be set in advance. These requirements limit the utility of a promising but non-significant trial once the maximum sample size is reached, because extending such a trial makes it impossible to control the type-I error. Moreover, also new trials cannot be added in a typical retrospective meta-analysis (not prespecified before any trial), 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 (Chapter 3). In contrast, an analysis based on E-variables can extend existing trials as well as inform whether to do new trials and meta-analyses, while still controlling type-I error rate. Type-I error control is retained even (i) if the e-value 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. Even with dependence between the trials, the test based on the multiplication of these e-values retains the type-I error control, as long as all trials test the same (i.e., global) null hypothesis. This becomes especially interesting if we want to combine the results of several trials in a bottom-up retrospective meta-analysis, where no top-down stopping rule can be enforced. We can even combine interim results of trials by multiplication while these trials are still ongoing – going beyond the realm of traditional sequential approaches.

Contributions and content We show that Cox' partial likelihood underlying his proportional hazards model can be used to define E-variables and test martingales. We first do this for (a) the case with only a group indicator (no other covariates) and without simultaneous events (ties) in Section 2.1.1, leading to a logrank test that is safe for optional stopping. We extend this to (b) the case with ties in Section 2.1.2 and to (c) a Gaussian approximation on the logrank statistic in Section 2.1.3 that is useful if only summmary statistics are available. We provide extensive computer simulations in Section 2.2 and Section 2.3, comparing our 'safe' logrank test to the traditional logrank test and α-spending approaches. In Section 2.2 we show that the exact safe 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 safe logrank test (case (c)). While always needing a bit more data in the design phase (the price for indefinite optional continuation), the expected sample size needed for rejection remains very competitive. We might want to design for a maximum sample size to achieve a certain power, but need a smaller sample size on average since we can safely engage in optional stopping. In Section 2.4 we extend the approach in various directions: first, in its basic version, the safe logrank test requires specification of a minimum clinically relevant effect size. If instead one wants to learn the actual effect size of the data and/or infuse prior knowledge about the effect size into the method via a Bayesian prior, this can be done without any difficulties. The resulting version of the safe test keeps providing non-asymptotic frequentist type-I error control even if these priors are wildly misspecified (i.e., they predict very different data from the data we actually observe); this is discussed in Section 2.4.1. We then show how our logrank test can be inverted to allow for *anytime-valid* confidence sequences (Section 2.4.2), and we provide the extension to covariates (Section 2.4.3). This extension, based on the Cox model, requires solving a complicated optimization problem and implementation is therefore deferred to future work.

To keep the exposition simple, when introducing our methods we represent data by a simplified discrete-time stochastic process, in which our test statistics take the form of likelihood ratios. In Appendix Section 2.A we show how our test statistics remain valid E-variables and test martingales under a proportional hazard assumption in continuous time; our likelihood ratios then become partial likelihood ratios. Once the definitions are in place, these results are mostly straightforward consequences from earlier work, in particular (Cox, 1972; Slud, 1992; Andersen et al., 1993). The novelty of our work is thus mainly in *defining* the new tests in the first place and showing by computer simulation that, while being substantially more flexible, they show competitive behavior with existing approaches, i.e. the classical logrank test in the fixed design setting and in combination with α-spending.

We delegate to the Appendix insights that, while important, are not needed to follow the main development. Most importantly, the particular E-variable we design satisfies the GROW criterion. Grünwald et al. (2019) provide several motivations for this criterion. We provide an additional one in Appendix Section 2.B by an argument (originally due to Breiman (1961), but not widely known) showing that it leads to tests with minimal expected stopping time. In the remainder of this introduction, we provide a short introduction to E-variables, test martingales and safe tests.

E-Variables, Test Martingales, Safety and Optimality

In this subsection we briefly introduce general concepts necessary to develop E-variables for the logrank test. The concepts are borrowed from Grünwald et al. (2019) (GHK from now on), which provides an extensive introduction to E-variables and its relation to likelihood ratios and Bayes factors. As seen in Example 1 below, when both the null H_0 and alternative hypotheses H_1 are simple, the most familiar E-variable is the likelihood ratio itself, where the product of likelihood ratios is also an E-variable. In fact, it is the best E-variable in the GROW sense defined further below. Similarly, when H_0 and or H_1 are composite, E-variables are often, but certainly not always, Bayes factors; and Bayes factors are certainly often, but not always E-variables. E-variables also have a crisp interpretation in terms of betting scores (GHK, Shafer (2021)). Briefly, a test martingale, which is the running product of a sequence of E-variables, describes the total profit you have made so far in a sequential gambling game in which you would not expect to win any money if the null were true. As a consequence, as expressed by (2.3) below, when the null holds true, there is little chance for the martingale to *ever* take on a large value. The general story that emerges from papers such as Shafer's as well as GHK and Ramdas et al. (2020) is that E-variables and test martingales are the 'right' generalization of likelihood ratios to the case that either or both H_0 and H_1 are composite. Existing tests based on generalizations of likelihood ratios to composite testing problems that are not E-variables often show problematic behavior in terms of nonasymptotic error control, whereas those generalizations that are E-variables can be combined freely over experiments while retaining type-I error control, thereby providing an intuitive notion of evidence.

Definition 2.0.1. *Let S be a nonnegative random variable defined on a sample space Ω, and let H_0, the null hypothesis, be a set of distributions for Ω. We call S an E-variable if for all $P \in H_0$, $\mathbf{E}_P[S] \leq 1$. For arbitrary random variable Z on Ω, S is called an E-variable* conditional on Z *if for all $P \in H_0$, $\mathbf{E}_P[S \mid Z] \leq 1$. A (conditional) E-variable is called* sharp *if for all $P \in H_0$, the inequality holds as an equality.*

We now consider not a fixed sample space, but a sequence of samples.

Definition 2.0.2. *Let $Y\langle 1\rangle, Y\langle 2\rangle, \ldots$ represent a discrete-time random process and let H_0, the null hypothesis, be a collection of distributions for this process. Fix $i > 0$ and let $S\langle i\rangle$ be a nonnegative random variable that is determined by (i.e. can be written as a function of) $(Y\langle 1\rangle, \ldots, Y\langle i\rangle)$. As an instance of Definition 2.0.1, for*[1] *$i \geq 0$, we say that $S\langle i+1\rangle$ is an E-variable conditionally on $(Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ if for all $P \in H_0$,*

$$\mathbf{E}_P\left[S\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle\right] \leq 1. \tag{2.1}$$

Definition 2.0.3. *If, for each i, $S\langle i\rangle$ is an E-variable conditional on $Y\langle 0\rangle, \ldots, Y\langle i-1\rangle$, then we say that the product process $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$ with $S^{\langle n\rangle} := \prod_{i=1}^{n} S\langle i\rangle$ is a* test supermartingale *relative to the given H_0. If all constituent E-variables are sharp, we call the process a test martingale.*

[1] For the case $i = 0$, (2.1) should be read as $\mathbf{E}_P\left[S\langle 1\rangle\right] \leq 1$.

We note that, for a supermartingale $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$, each $S^{\langle i\rangle}$ is itself an (unconditional) E-variable.[2]

Example 1. [(Partial) Likelihood Ratios as E-Values and Test Martingales] Let $H_0 = \{P_0\}$ and $H_1 = \{P_1\}$ both be simple, each containing a single distribution for the process $Y\langle 1\rangle, Y\langle 2\rangle, \ldots$ with each $Y\langle i\rangle$ taking values in a finite set $\mathscr{Y}$. Let, for each $i \in \mathbb{N}$, $p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ and $p_1(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ be the conditional probability mass functions corresponding to P_0 and P_1. Then $S\langle i+1\rangle := p_1(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)/p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)$ is a sharp E-variable, as is seen from calculating (2.1) (an explicit calculation is given for the logrank test in (2.9)). Further, $S^{\langle n\rangle} = \prod_{i=1}^{n} p_1(Y\langle i\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i-1\rangle)/p_0(Y\langle 1\rangle, \ldots, Y\langle i\rangle \mid Y\langle i-1\rangle) = p_1(Y\langle 1\rangle, \ldots, Y\langle n\rangle)/p_0(Y\langle 1\rangle, \ldots, Y\langle n\rangle)$ is the likelihood ratio for the first n outcomes, and the test supermartingale $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$ is a likelihood ratio process.

More generally, let $Y\langle 1\rangle, Y\langle 2\rangle, \ldots$ be a discrete stochastic process defined on an arbitrary underlying measurable space (which may e.g. represent time as a continuous-valued random variable) and let H_0 be any set of probability distributions on this space such that all elements in H_0 agree on the conditional probability mass functions $p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle))$ for $i = 1, 2 \ldots$. Let $q(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)), i = 1, 2, \ldots$ and $r(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)), i = 1, 2, \ldots$ denote any other sequence of conditional probability mass functions for this discrete process. Then we have, by the same explicit calculation, that

$$S\langle i+1\rangle := \frac{r(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)}{q(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)} \text{ is a sharp conditional } E\text{-variable}$$
$$\text{if } p_0(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)) = q(Y\langle i+1\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i\rangle)). \tag{2.2}$$

If this is indeed the case for each i, then $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \ldots$ with $S^{\langle n\rangle} = \prod_{i=1}^{n} r(Y\langle i\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i-1\rangle)/p_0(Y\langle i\rangle \mid Y\langle 1\rangle, \ldots, Y\langle i-1\rangle) =$
$r(Y\langle 1\rangle, \ldots, Y\langle n\rangle)/p_0(Y\langle 1\rangle, \ldots, Y\langle n\rangle)$ is a test martingale and $S^{\langle n\rangle}$ is now in general not a full, but so-called *partial* likelihood — but the difference will not matter for our purposes.

For the special case of likelihood ratios, the following safety result is referred to as a *universal bound* by Royall (1997).

Safety The interest in E-variables and test martingales derives from the fact that we have type-I error control irrespective of the stopping rule used: for any test (super-) martingale $\{S^{\langle i\rangle}\}_{i\in\mathbb{N}}$ relative to $\{Y\langle i\rangle\}_{i\in\mathbb{N}}$ and H_0, Ville's inequality (Ville, 1939; Shafer et al., 2011) tells us that for all $0 < \alpha \leq 1$, $P \in H_0$,

$$P(\text{there exists } i \text{ such that } S^{\langle i\rangle} \geq 1/\alpha) \leq \alpha. \tag{2.3}$$

In other words, under the null there is little chance (e.g., less than or equal to 5%) that a supermartingale $S^{\langle i\rangle}$ will *ever* realize a large value (e.g., larger than 20). Ville's inequality

[2] Readers familiar with measure-theoretic terminology may recognize a test (super-) martingale as a non-negative (super-) martingale relative to the filtration $\sigma(Y\langle 1\rangle), \sigma(Y\langle 1\rangle, Y\langle 2\rangle), \ldots$. The requirement that '$S\langle i\rangle$ is determined by $(Y\langle 1\rangle, \ldots, Y\langle i\rangle)$' then means that the process $\{S\langle i\rangle\}_{i\in\mathbb{N}}$ is adapted to this filtration.

implies that type-I error control is guaranteed regardless of the stopping rule, which may be deterministic, e.g., stop after 100 samples, or data-driven, e.g., stop as soon as $S^{\langle i \rangle} \geq 20$, or may be exogenuously determined, e.g. stop when your boss tells you to.

Thus, if we measure evidence against the null hypothesis after observing i data units by $S^{\langle i \rangle}$, and we reject the null hypothesis if $S^{\langle i \rangle} \geq 1/\alpha$, then our type-I error will be bounded by α, no matter what stopping rule was used for determining i. We thus have type-I error control if we run out of data, or money to gather new data, and even if we use the most aggressive stopping rule compatible with this scenario: stop at the first i at which $S^{\langle i \rangle} \geq 1/\alpha$. We also have type-I error control if the actual stopping rule is unknown to us, or determined by external factors independent of the data $Y\langle i \rangle$ – as long as the decision whether to stop depends only on past data, and not on the future. This is impossible to achieve with standard Neyman-Pearson tests.

We will call any procedure that takes as input stopping time τ, significance level α and the sequence of random variables $S\langle 1 \rangle, S\langle 2 \rangle, \ldots$, that stops at (random) time τ and that outputs REJECT if $S^{\langle \tau \rangle} \geq 1/\alpha$ and $accept$ otherwise, a *level α-test that is safe under optional stopping*, or simply a *safe test*.

Importantly, we can also deal with *optional continuation*: we can combine E-variables from different trials that share a common null (but may be defined relative to a different alternative) by multiplication, and still retain type-I error control – see Example 3. If we used p-values rather than E-variables we would have to resort to e.g. Fisher's method, which, in contrast to multiplication of e-values, is invalid if there is a dependency between the (decision to perform) tests. Finally, E-variables and test martingales can also be used to define 'anytime-valid confidence intervals' that remain valid under optional stopping (Section 2.4.2).

Optimality Just like p-values, E-variables only require the specification of a null model H_0. Provided with such a null model, we can typically specify a large class of E-variables, denoted here by $\mathscr{S}$, all which remain small with high probability under the null. If, on the other hand, the data are governed by a distribution belonging to a given alternative model H_1, then we would like to choose that E-variable from $\mathscr{S}$ that accumulates evidence against the null as fast as possible. The speed of evidence accumulation under the alternative is defined (conservatively) as the smallest expectation of the logarithm of the E-variable under the alternative, see GHK and Shafer (2021) for various reasons for this choice. More specifically, provided with an alternative H_1, the growth rate of a conditional E-variable $S\langle i \rangle$ under distribution $P_1 \in H_1$ is defined as $\mathbf{E}_{P_1}[\log S\langle i+1 \rangle \mid Y\langle 1 \rangle, \ldots, Y\langle i \rangle]$. The optimal E-variable amongst all E-variables conditional on $Y\langle 1 \rangle, \ldots, Y\langle i \rangle$ is that S that solves the criterion

$$\max_{S} \min_{P \in H_1} \mathbf{E}_P[\log S\langle i+1 \rangle \mid Y\langle 1 \rangle, \ldots, Y\langle i \rangle]. \tag{2.4}$$

We call the optimal E-variable GROW, i.e., growth-rate optimal in worst-case. Appendix Section 2.B provides another motivation for using the logarithm to measure the speed of evidence accumulation, originally provided by Breiman (1961). Specifically, we show

that, under the alternative, the GROW E-variable minimizes the expected number of data points needed to reject the null if no maximum sample size is specified. Note that this is analogous to finding a test that maximizes power. In Section 2.3 we provide some simulations to relate power to GROW. Note that we cannot directly use power in designing tests, since the notion of power requires a fixed sampling plan, which by design we do not have.

2.1 Safe logrank tests

The classical logrank test compares the risk of an event in two groups (e.g. a treatment and a control group) in a nonparametric test, meaning that it requires no underlying (parameterized) distribution on the event times. No matter when the events occur, the null hypothesis assumes that their probability is equal for all participants, regardless of their group. We can flip this null hypothesis on its head for each event time: given that an event has occurred, it has equal probability to have been in either the treatment or the control group, when an equal number of participants are at risk in both groups; more generally, the probability ratio between an event in both groups is equal to the ratio of participants in these groups. That risk set (of participants at risk) changes with each event so the test crucially relies on the order (the ranks) of the events.

Logrank hypotheses tested We observe a sequence of event times $t\langle 1\rangle < t\langle 2\rangle < t\langle 3\rangle < \dots$ such that for all i, at time $t\langle i\rangle$, one event happens, and inbetween $t\langle i\rangle$ and $t\langle i+1\rangle$, no events happen. The time until the i^{th} event occurs – $t\langle i\rangle$ itself – does not play a role in the logrank statistic; only the ordering of events matters. The hypothesis tested by the logrank test can be expressed in terms of the instantaneous risk to experience an event – the hazard λ_1 in the treatment group and λ_0 in the control group – at each event time $t\langle i\rangle$ (Klein and Moeschberger, 2006, p. 206):

$$\begin{aligned} H_0: &\quad \lambda_1(t\langle i\rangle) = \lambda_0(t\langle i\rangle), && \text{for all event times } i,\\ H_1: &\quad \lambda_1(t\langle i\rangle) \text{ and } \lambda_0(t\langle i\rangle) \text{ are different}, && \text{for some event times } i. \end{aligned}$$

Logrank statistic for single events Let $Y_1\langle i\rangle$ denote that number of participants in the risk set that are in the treatment group at the time of the i^{th} event and $Y_0\langle i\rangle$ for the number of participants at risk in the control group. We use the standard notational convention that $Y_1\langle i\rangle + Y_0\langle i\rangle$ includes the participant experiencing the i^{th} event. Let $O_1\langle i\rangle$ and $O_0\langle i\rangle$ count the number of observed events in the treatment group and control group at the i^{th} event time. These always describe a single event in this section: $O_1\langle i\rangle = 1$ and $O_0\langle i\rangle = 0$ if the event occurred in the treatment group, and $O_1\langle i\rangle = 0$ and $O_0\langle i\rangle = 1$ if a single event occurred in the control group. We extend this Bernoulli case to multiple simultaneous events (ties) – in which case $O_1\langle i\rangle$ can be larger than 1 – in Section 2.1.2. The logrank statistic for a sample size of n single events with nonempty risk sets (i.e.

$Y_1\langle i\rangle > 0, Y_0\langle i\rangle > 0$) is the following (for the treatment group 1):

$$Z = \frac{\sum_{i=1}^{n}\{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_{i=1}^{n} V_1\langle i\rangle}} \quad \text{with} \quad E_1\langle i\rangle = \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle}; \quad V_1\langle i\rangle = E_1\langle i\rangle \cdot \big(1 - E_1\langle i\rangle\big). \tag{2.5}$$

For sufficiently large sample size (number of events n), the logrank Z-statistic has an approximate standard normal (Gaussian) distribution under the null hypothesis.

2.1.1 The safe logrank test for single events

Logrank partial likelihood for single events If we assume that our hazards follow the Cox (1972) proportional hazards model, we obtain a more explicit alternative hypothesis for the logrank test in terms of a constant hazard ratio θ:

$$\begin{aligned} H_0: &\quad \lambda_1(t\langle i\rangle) = \theta \cdot \lambda_0(t\langle i\rangle), && \text{for all event times } t\langle i\rangle, \text{ with } \theta = 1, \\ H_1: &\quad \lambda_1(t\langle i\rangle) = \theta \cdot \lambda_0(t\langle i\rangle), && \text{for all event times } t\langle i\rangle, \text{ with } \theta \neq 1. \end{aligned}$$

This turns the nonparametric logrank test into a semi-parametric test for which we can formulate a partial likelihood. For now, we model the data by a simplified process in which this partial likelihood is simply a standard likelihood, returning to the 'partial' interpretation further below.

We define a risk set process $\vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \vec{Y}\langle 3\rangle, \ldots$ with $\vec{Y}\langle i\rangle = \big(Y_0\langle i\rangle, Y_1\langle i\rangle\big)$ that describes how many participants are at risk at each event time in the treatment and control group. For simplicity, we assume no censoring, but the likelihood we develop remains valid if the risk set changes because of left-truncation or noninformative right-censoring. At the time of the first event, everyone is at risk, captured by $Y_1\langle 1\rangle = m_1$, the number of participants allocated to the treatment group, and $Y_0\langle 1\rangle = m_0$, the number allocated to the control group. At each event time i, if $O_1\langle i\rangle = 1$ (event in treatment group) then $Y_1\langle i+1\rangle := Y_1\langle i\rangle - 1$ and $Y_0\langle i+1\rangle = Y_0\langle i\rangle$, and analogously for the control group. To complete the specification of the process, we now specify the probability that the single event occurs in the treatment group, given that we have reached a new event time i. This probability only depends on the risk set at the i^{th} event time, so we set:

$$\begin{aligned} &P_\theta(O_1\langle i\rangle = o_1 \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle i-1\rangle, \vec{Y}\langle i\rangle) := P_\theta(O_1\langle i\rangle = o_1 \mid \vec{Y}\langle i\rangle); \\ &P_\theta(O_1\langle i\rangle = o_1 \mid \vec{Y}\langle i\rangle = (y_0, y_1)) := q_\theta(o_1 \mid (y_0, y_1)), \end{aligned}$$

where $o_1 \in \{0, 1\}$ and q_θ is the conditional probability mass function of a treatment group event, given that the i^{th} event occurred. That is,

$$q_\theta\big(o_1 \mid (y_1, y_0)\big) = \left(\frac{y_1 \cdot \theta}{y_0 + y_1 \cdot \theta}\right)^{o_1} \left(\frac{y_0}{y_0 + y_1 \cdot \theta}\right)^{1-o_1} \tag{2.6}$$

is the probability mass function of a Bernoulli $y_1\theta/(y_0 + y_1\theta)$-distribution. This really expresses that, given that there is an event, the probability of being the participant with this event is $\theta/(y_0 + y_1\theta)$ for each participant in the treatment group, and $1/(y_0 + y_1\theta)$ for each participant in the control group. Summing the probabilities gives (2.6).

Our process is fully specified by the product of conditional probability mass functions (2.6). In particular, at the sample size of n observed events, the joint likelihood (probability mass of these n events as a function of the parameter θ conditional on the data) is given by

$$\mathfrak{L}\big(\theta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle\big) = \prod_{i=1}^{n} q_\theta\big(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)\big). \tag{2.7}$$

We can think of (2.7) as a standard likelihood in the sample space implicitly defined above, giving an extremely simplified handling of time. As we shall see in Section 2.4.3, we can also think of it as a special case of Cox' partial likelihood, obtained if there are no covariates except for the treatment/control indicator. Further below we shall motivate it further in terms of continuous time. Below, we illustrate the connection to the classical logrank test.

Logrank score test for single events If we define $\beta = \log\theta$ (following Cox (1972)), take the score function $U(\beta)$ of the likelihood in (2.7) and evaluate at $\beta = 0$, we get the familiar ingredients of logrank statistic in terms of observed and expected events:

$$U(\beta) = \frac{\mathrm{d}\log \mathfrak{L}(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta} = \sum_{i=1}^{n}\left\{O_1\langle i\rangle - \frac{Y_1\langle i\rangle \cdot \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)}\right\},$$

$$-U'(\beta) = -\frac{\mathrm{d}^2 \log \mathfrak{L}(\beta \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta^2} = \sum_{i=1}^{n}\left\{\frac{Y_1\langle i\rangle \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)}\frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)}\right\},$$

$$U(0) = \sum_{i=1}^{n}\left\{O_1\langle i\rangle - \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle}\right\} = \sum_{i=1}^{n}\big\{O_1\langle i\rangle - E_1\langle i\rangle\big\},\ -U'(0) = \sum_{i=1}^{n} E_1\langle i\rangle\big(1 - E_1\langle i\rangle\big).$$

Under the null hypothesis, standardizing $U(0)$ with the variance $-U'(0)$ (dividing by the observed Fisher information) gives the logrank Z-statistic in (2.5). This shows that the logrank test is the score test corresponding to the likelihood (2.7), a fact already expressed by Cox (1972) when introducing the proportional hazards model. A more detailed derivation is provided in Appendix Section 2.C.

Logrank E-variable For given $\theta_0, \theta_1 > 0$, define the following one-outcome likelihood ratio

$$M_{\theta_1,\theta_0}\langle i\rangle := \frac{\mathfrak{L}\big(\theta_1 \mid \vec{Y}\langle i\rangle\big)}{\mathfrak{L}\big(\theta_0 \mid \vec{Y}\langle i\rangle\big)} = \frac{q_{\theta_1}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)}{q_{\theta_0}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)}. \tag{2.8}$$

Since the likelihood ratio of the i^{th} event time depends only on the risk set at the i^{th} event time $\vec{Y}\langle i\rangle$, we can write out the expectation as follows

$$\begin{aligned}\mathbf{E}_{P_{\theta_0}}\big[M_{\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle\big] &= \mathbf{E}_{P_{\theta_0}}\big[M_{\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle i\rangle = (y_1, y_0)\big] \\ &= \sum_{o_1\in\{0,1\}} q_{\theta_0}(o_1 \mid y_1, y_0)\cdot\frac{q_{\theta_1}(o_1 \mid y_1, y_0)}{q_{\theta_0}(o_1 \mid y_1, y_0)} = \sum_{o_1\in\{0,1\}} q_{\theta_1}(o_1 \mid y_1, y_0) = 1.\end{aligned} \tag{2.9}$$

This standard argument (as in Example 1) immediately shows that, under P_{θ_0}, for all i and all $\theta_1 > 0$, $M_{\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on $\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$, and $M^{\langle 1\rangle}_{\theta_1,\theta_0}, M^{\langle 2\rangle}_{\theta_1,\theta_0}, \ldots$ with

$$M^{\langle n\rangle}_{\theta_1,\theta_0} := \prod_{i=1}^{n} M_{\theta_1,\theta_0}\langle i\rangle \tag{2.10}$$

is a test martingale under P_{θ_0} relative to the risk set process $\vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \vec{Y}\langle 3\rangle, \ldots$. Here we note that $M_{\theta_1,\theta_0}\langle i\rangle$ corresponds to $S\langle i+1\rangle$ in Definition 2.0.2, and similarly $M^{\langle i\rangle}_{\theta_1,\theta_0}$ corresponds to $S^{\langle i+1\rangle}$, reflecting existing different notational conventions in the test martingale and survival analysis literature.

Thus, by Ville's inequality, we have the desired:

$$P_{\theta_0}\Big(\text{there exists } n \text{ with } M^{\langle n\rangle}_{\theta_1,\theta_0} \geq 1/\alpha\Big) \leq \alpha. \tag{2.11}$$

We can generalize M_{θ_1,θ_0} by replacing q_{θ_1} in (2.8) by another conditional probability mass function $r_i(\cdot \mid \ldots)$ on $x \in \{0,1\}$, allowed to depend on i and the full past sequence of risk sets $Y\langle 1\rangle, \ldots, Y\langle i\rangle$. For any given sequence of such conditional probability mass functions, $r \equiv \{r_i\}_{i\in\mathbb{N}}$, we extend definition (2.8) to

$$M_{r,\theta_0}\langle i\rangle = \frac{r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle)}{q_{\theta_0}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)} \quad ; \quad M^{\langle n\rangle}_{r,\theta_0} := \prod_{i=1}^{n} M_{r,\theta_0}\langle i\rangle. \tag{2.12}$$

For any choice of the r_i, the analogue of (2.9) clearly still holds for the resulting M_{r,θ_0}, making $M_{r,\theta_0}\langle i\rangle$ a conditional E-variable and its product process a martingale; and then Ville's inequality (2.11) must also still hold.

The Underlying Continuous Time Model The logrank test is usually justified based on an underlying continuous time model. We now show that tests using $M^{\langle n\rangle}_{r,\theta_0}$ remain valid under this same underlying model, which we now describe. We identify each of the $m = m_0 + m_1$ participants by an index $j \in \{1, \ldots, m\}$. The vector $\vec{g} \in \{0,1\}^m$ encodes group assignment, $\vec{g}_j$ denoting the group assignment of participant j, so that the number of 0-components of $\vec{g}$ is m_0. We assume that the group assignment is fixed in advance. For simplicity we assume no censoring, but as said, all results remain valid if the risk set changes because of left-truncation or noninformative right-censoring. We assume that each participant $j \in \{1, \ldots, m\}$ has an event at time T_j, where $T_1, T_2, \ldots, T_m$ are independent, and T_j has distribution that is fully determined by j's group membership: T_j has distribution $P_{\vec{g}_j}$ if patient j is in group g. For each participant j we let $S_j(t) = P_{\vec{g}_j}\{T_j > t\}$ be its survival function and let $\lambda_{\vec{g}_j}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\ln S_i(t)$ its hazard function, which we assume exists and is continuous. It will then depend on the participant only through its group membership. The relation between the survival function and the hazard function implies that the conditional probability that the event time T_j falls in the interval $(t, t+h]$ given that $T_j > t$, can be computed as

$$P_{\vec{g}_j}\{t < T_j \leq t+h \mid T_j > t\} = 1 - \exp\Big(\int_t^{t+h} \lambda_{\vec{g}_j}(s)\,\mathrm{d}s\Big) \tag{2.13}$$

for any $t > 0$. Under the proportional hazards ratio model, the ratio $\lambda_1(t)/\lambda_0(t)$ remains constant and takes the value θ. The results of Slud (1992) imply that under these assumptions (i.e. independence of T_j and proportional hazards), the conditional distribution of $O_1\langle i\rangle$ given $Y_1\langle i\rangle, Y_0\langle i\rangle$ is indeed uniquely defined and given by $q_{\theta_0}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)$ (see also Andersen et al. (1993) for closely related results); for convenience we give an informal derivation (avoiding measure theory and ignoring censoring) of this result in Appendix Section 2.A.

Combining this result with (2.2) we see that $M^{\langle n\rangle}_{r,\theta_0}$ remains a test martingale within this standard continuous time model, and Ville's inequality remains valid.

Example 2. [GROW alternative] The simplest possible scenario is that of a one-sided test between the 'no effect' null hypothesis H_0 ($\theta_0 = 1$) and a one-sided alternative hypothesis $H_1 = \{P_{\theta_1} : \theta_1 \in \Theta_1\}$ represented by a minimal clinically relevant effect size $\theta_{\min}$. For example, if 'event' means that the participant gets ill, then we would hope that under the treatment, $\theta_{\min}$ would be a value smaller than 1 and we would have $\Theta_1 = \{\theta_1 : 0 < \theta_1 \leq \theta_{\min}\}$. If 'event' means 'cured' then we would typically set $\Theta_1 = \{\theta_1 : \theta_{\min} \leq \theta_1 < \infty\}$ for some $\theta_{\min} > 1$. We will take the left-sided alternative with $\theta_{\min} < 1$ as a running example, but everything we say in the remainder of this chapter also holds for the right-sided alternative. The GROW (*growth-optimal in worst-case*) E-variable as in (2.4) is given by taking $M_{\theta_{\min},\theta_0}$, i.e. it takes the $\theta_1 \in \Theta_1$ closest to θ_0. That is,

$$\max_{\theta_1>0} \min_{\theta\in\Theta_1} \mathbf{E}_{P_\theta}[\log M_{\theta_1,\theta_0}\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle] = \max_{r} \min_{\theta\in\Theta_1} \mathbf{E}_{P_\theta}[\log M_{r,\theta_0}\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle]$$

is achieved by setting $\theta_1 = \theta_{\min}$, no matter the values taken by $Y_1\langle i\rangle, Y_0\langle i\rangle$. Here the second maximum is over all sequences of conditional distributions r_i as used in (2.12). Thus, among all E-variables of the general form $M_{r,\theta_0}\langle i\rangle$ there are strong reasons for setting $r_i = q_{\theta_{\min}}$ – this is further elaborated in Section 2.4 and Appendix Section 2.B.

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

$$M^{\langle i\rangle}_{\text{two-sided}} := \frac{1}{2}\left(M^{\langle i\rangle}_{\theta_{\min},\theta_0} + M^{\langle i\rangle}_{1/\theta_{\min},\theta_0}\right). \tag{2.14}$$

It is then easy to verify that for each i, $M_{\text{two-sided}}\langle i\rangle := M^{\langle i\rangle}_{\text{two-sided}}/M^{\langle i-1\rangle}_{\text{two-sided}}$ is a conditional E-variable, i.e. $\mathbf{E}_{P_{\theta_0}}\left[M_{\theta_1,\theta_0}\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle\right] = 1$, and $M_{\text{two-sided}}\langle 1\rangle, M_{\text{two-sided}}\langle 2\rangle, \ldots$ constitutes a test martingale; see GHK for details.

Example 3. [Meta-analysis] What if we want to combine several trials, conducted in different hospitals or in different countries? In such a case we often compare a 'global' null – H_0 is true in all trials – to an alternative that allows for different hazard ratios in different trials, with different populations. We may thus associate the i^{th} event at the k^{th} trial with E-variable $M_{\theta_{1,k},\theta_0}\langle i,k\rangle$, with $\theta_{1,k}$ varying from trial to trial.

$$M\langle i,k\rangle := \frac{q_{\theta_{1,k}}(O_1\langle i,k\rangle \mid Y_1\langle i,k\rangle, Y_0\langle i,k\rangle)}{q_{\theta_0}(O_1\langle i,k\rangle \mid Y_1\langle i,k\rangle, Y_0\langle i,k\rangle)}$$

denotes the E-variable corresponding to the i^{th} event in the k^{th} trial, with $Y_1\langle i,k\rangle$ and $Y_0\langle i,k\rangle$ denoting the number of people at risk in the treatment and control group of trial k at the i^{th} event. The evidence against H_0 after having observed n_k events from trial k, for $k \in \mathscr{K}$, with $\mathscr{K}$ the subset of all trials for which some data is already available can then be summarized as

$$M_{\text{META}} := \prod_{k\in\mathscr{K}} \prod_{i=1,\ldots,n_k} M\langle i,k\rangle.$$

As GHK explain, in such cases the anytime-valid type-I error guarantee still holds: under the global null, where $\theta = \theta_0$ in all trials, the probability that there ever comes a sequence of events in any combination of trials such that for this sequence, $M_{\text{META}} \geq 1/\alpha$, is still bounded by α. Thus, we effectively perform an *on-line, cumulative* and possibly *live* meta-analysis here that remains valid irrespective of the order in which the events of the different trials come in. Importantly, unlike in α-spending approaches, the maximum number of trials and the maximum sample size (number of events) per trial do not have to be fixed in advance; we can always decide to start a new trial, or to postpone to end a trial and wait for additional events. This has many advantages in terms of collaboartion, efficiency and communication of results (see Chapter 1).

2.1.2 Allowing for ties

In many settings we may observe ties: we cannot identify distinct event times for all observed events and consider some of those events to have happened simultaneously. To formalize this we first define $Y\langle i\rangle = Y_1\langle i\rangle + Y_0\langle i\rangle$ and $O\langle i\rangle = O_1\langle i\rangle + O_0\langle i\rangle$. In case of a tie at event time i, $O\langle i\rangle$ is larger than one and consists of multiple events in the treatment group ($O_1\langle i\rangle \geq 1$) and/or in the control group ($O_0\langle i\rangle \geq 1$). We cannot represent this common situation with our simple process P_θ from Section 2.1.1, since it requires a fully observable ordering of events. But we can easily extend it to ties by conditioning on the number of observed events $O\langle i\rangle = o$ at each of the event times i with multiple events. The probability distribution of the number of events in the treatment group $O_1\langle i\rangle$ is defined by the Fisher noncentral hypergeometric distribution. Here $O_1\langle i\rangle = o_1$ indicates the number of events that are observed in the treatment group, out of a total of $O\langle i\rangle = o$ events (so $O_1\langle i\rangle$ can be $0, 1, \ldots, o$). The distribution q_θ defined below replaces q_θ from (2.7):

$$q_\theta(o_1 \mid (y_0,y_1),o) := \frac{\binom{y_1}{o_1}\cdot\binom{y_0}{o-o_1}\cdot\theta^{o_1}}{\sum_{u=o_1^{\min}}^{o_1^{\max}}\binom{y_1}{u}\cdot\binom{y_0}{o-u}\cdot\theta^{u}} \quad \text{with} \quad \begin{cases} o_1^{\min} = \max\{0, o-y_0\} \\ o_1^{\max} = \min\{o, y_1\}. \end{cases} \tag{2.15}$$

This is the probability mass function of a Fisher noncentral hypergeometric distribution with parameters $(O\langle i\rangle, Y_1\langle i\rangle, Y\langle i\rangle, \theta) = (o, y_1, y, \theta)$. In the Appendix Section 2.C we show that also for this partial likelihood, the score test equals the logrank test. When dealing with ties we must distinguish between the number of events n in a sample and the number of event times $I \leq n$ in the same sample. The corresponding martingale $M^{\langle I\rangle}_{\theta_1,\theta_0} = \prod_{i=1}^{I} M_{\theta_1,\theta_0}\langle i\rangle$ is now a product of I E-variables, together covering n events. In Appendix Section 2.A we show that, in the continuous-time model, $M^{\langle I\rangle}_{\theta_1,\theta_0}\langle i\rangle$ is still a conditional E-variable, and $M^{\langle I\rangle}_{\theta_1,\theta_0}$ is a test martingale *if the null* $\theta_0 = 1$, i.e. under the

assumption that there are no differences between the two groups; for other θ_0, it is only an 'approximate' E-variable, becoming more exact the closer the time points at which we check whether event(s) have happened lie together in continuous time. Thus, we have a weaker result than for the case without ties, for which M_{θ_1,θ_0} is a test martingale in the continuous time setting under arbitrary θ_0 — but as long as we test with null $\theta_0 = 1$, all our results are still exact. Again, for simplicity we ignore censoring in the appendix. The present approach for ties can still be adapted to noninformative right censoring under the additional common assumption that the events reported at each observation time precede any censorings, so that censored patients contribute fully to the risk sets under consideration.

Compatibility between q_θ in (2.15) and (2.6) Reassuringly, if at all event times $i = 1, 2, \ldots, I$ we observe only a single event, q_θ from (2.15) and q_θ from (2.7) are the same, since for a single event in the treatment group ($o = 1, o_1 = 1$) and a given $y_1, y_0 > 1$, we get

$$q_\theta(1 \mid (y_0, y_1), 1) = \frac{\binom{y_1}{1} \cdot \binom{y_0}{0} \cdot \theta^1}{\sum_{v=0}^{1} \binom{y_1}{u} \cdot \binom{y_0}{1-u} \cdot \theta^u} = \frac{y_1 \cdot \theta}{y_0 + y_1 \cdot \theta},$$

and analogously for a single event in the control group ($o = 1, o_1 = 0$).

2.1.3 Gaussian approximation on the logrank statistic

Our GROW safe logrank test is an exact logrank test for a risk set process with a hypergeometric probability mass function (2.15) under the null hypothesis ($\theta = 1$) – with the Bernoulli probability mass function (2.6) as a special case for single events. The exact test can be used for a survival data set of event and censoring times that captures the full risk set process. Here we describe an approximation to this safe logrank test based on a sequential-Gaussian approximation on the logrank statistic. The approximation is of interest for two reasons. First, in practical situations, sometimes only the logrank Z-statistic and other summary statistics are available, and not the full risk set process. If we also know the number of events n and initial number of participants in the two groups m_1 and m_0, the Gaussian approximation can then still be used. Second, the group sequential and α-spending approaches that we compare ourselves to in the next section are based on a Gaussian approximation to the logrank statistic. The behavior of the Gaussian approximation can (and will) give us insights into how the safe logrank test compares to the group sequential and α-spending approaches as well.

For a mix of single and tied events, the logrank Z-statistic for n events at I event times is defined as follows:

$$Z = \frac{\sum_{i=1}^{I} \{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_{i=1}^{I} V_1\langle i\rangle}} \quad \text{where} \tag{2.16}$$

$$E_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle; \qquad A_1\langle i\rangle = \frac{Y_1\langle i\rangle}{Y\langle i\rangle}; \qquad V_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle \cdot \big(1 - A_1\langle i\rangle\big) \cdot \frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}.$$

For single events, $O\langle i\rangle = 1$ and $E_1\langle i\rangle = A_1\langle i\rangle$. If all event times have single events, then $n = I$. In general, with and without ties, $n = \sum_{i=1}^{I} O\langle i\rangle$. The above formulation is also found in Cox (1972, equation (26)) and $\frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}$ is described as a "multiplicity" correction or "correction for ties" (Klein and Moeschberger, 2006, p. 207).

Under the null hypothesis, this logrank statistic has an approximate standard normal (Gaussian) distribution ($Z \sim \mathcal{N}(0,1)$). Under the alternative distribution, Schoenfeld (1981) gives an asymptotic result for survival data (no ties) following the proportional hazards model with hazard ratio θ: a Taylor approximation around $\theta = 1$ gives that the logrank Z-statistic is also normally distributed ($Z \sim \mathcal{N}(\mu, 1)$) with:

$$\mu \approx \frac{\sum_{i=1}^{n} \log(\theta) E_1\langle i\rangle \big(1 - E_1\langle i\rangle\big)}{\sqrt{\sum_{i=1}^{n} E_1\langle i\rangle \big(1 - E_1\langle i\rangle\big)}} \approx \log(\theta)\sqrt{nE_1\langle 1\rangle\big(1 - E_1\langle 1\rangle\big)} = \log(\theta)\sqrt{\frac{m_1 \cdot m_0}{(m_1 + m_0)^2}}\sqrt{n}. \tag{2.17}$$

Schoenfeld's assumptions Schoenfeld's asymptotic result heavily relies on two properties: (a) the mean of the alternative is close enough to one so that the first-order Taylor approximation around $\theta = 1$ is adequate and (b) $E_1\langle i\rangle$ stays approximately constant at all event times i, i.e. close to the initial allocation proportion $E_1\langle 1\rangle = m_1/(m_1 + m_0)$. These two properties indicate that this asymptotic distribution is only reasonably good if the hazard ratio θ is close to 1 and the initial risk set m_0 and m_1 are both large in comparison to the number of events and the amount of censoring, in which case also the multiplicity correction in the definition of $V_1\langle i\rangle$ for ties is negligible.

Logrank statistic per event time This raises the question whether a Gaussian approximation is sensible for a logrank statistic per event time i: a priori it is not at all clear whether Schoenfeld's asymptotic, fixed sample result has a nonasymptotic sequential counterpart. We define the logrank statistic per event time

$$Z\langle i\rangle = \frac{O_1\langle i\rangle - E_1\langle i\rangle}{\sqrt{V_1\langle i\rangle}}. \tag{2.18}$$

and check whether the exact E-variable (with the q for a mix of ties and single events from (2.15))

$$M_{\theta_1,\theta_0}\langle i\rangle = \frac{q_{\theta_1}\big(O_1\langle i\rangle \mid (Y_1\langle i\rangle, Y_0\langle i\rangle), O\langle i\rangle\big)}{q_{\theta_0}\big(O_1\langle i\rangle \mid (Y_1\langle i\rangle, Y_0\langle i\rangle), O\langle i\rangle\big)} \text{ behaves similar to } M'_{\mu_1,\mu_0}\langle i\rangle := \frac{\phi_{\mu_1\sqrt{O\langle i\rangle}}(Z\langle i\rangle)}{\phi_{\mu_0}(Z\langle i\rangle)} \tag{2.19}$$

for $\theta_0 = 1, \mu_0 = 0$ and $\mu_1 = \log(\theta_1) \cdot \sqrt{(m_1 \cdot m_0)/(m_1 + m_0)^2}$, where ϕ_μ is the Gaussian density with mean μ and variance 1.

Safety only for balanced allocation We henceforth focus on the case $\theta_0 = 1$, such that that $\mu_0 = 0$. Figure 2.1 shows that in case of balanced 1:1 allocation $M'_{\mu_1,0}\langle i\rangle$ is an E-variable, since its expectation is 1 or smaller. However, in case of unbalanced 2:1 or

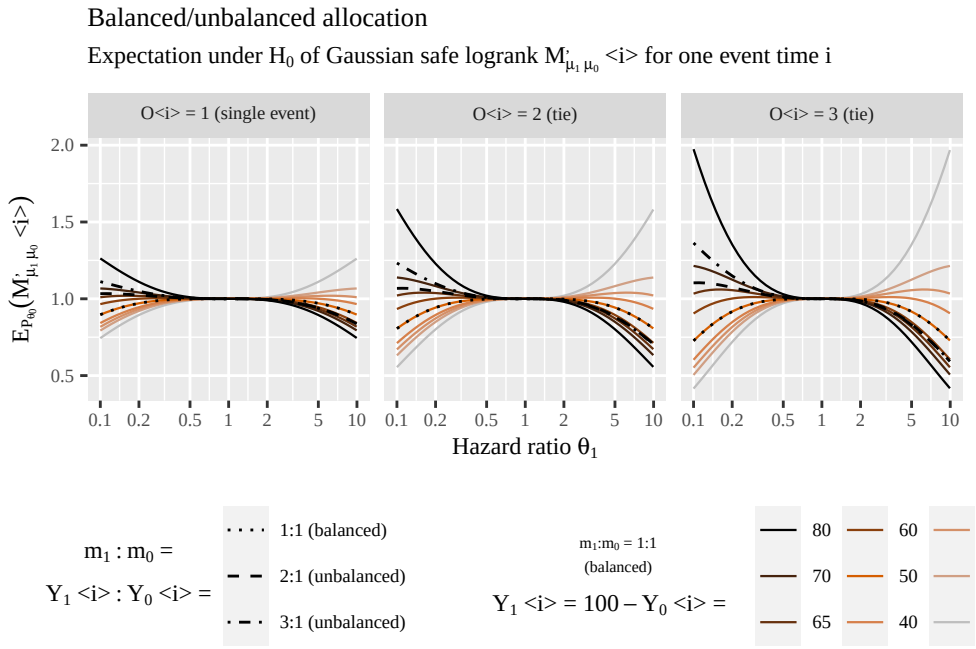

Figure 2.1. *For balanced allocation $M'_{\mu_1,\mu_0}\langle i\rangle$ is an E-variable, but it is not for unbalanced allocation. The risk set can also start out balanced but become unbalanced, in which case $M'_{\mu_1,\mu_0}\langle i\rangle$ is also not an E-variable. Here $\mu_0 = 0$, and μ_1 follows from θ_1 as in (2.19); in the expectation $E_{P_{\theta_0}}$ under the null hypothesis $\theta_0 = 1$. Note that for $i > 1$, $M'_{\mu_1,\mu_0}\langle i\rangle$ depends both on the present risk set balance $Y_1\langle i\rangle : Y_0\langle i\rangle$ and on the initial balance $m_1 : m_0$. Note also that the x-axis is logarithmic.*

3:1 allocation and designs with hazard ratio $\theta_1 < 1$, $M'_{\mu_1,0}\langle i\rangle$ is not an E-variable. Of course, the risk set can also start out balanced by allocation, but become unbalanced. Figure 2.1 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 recommend to not 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 as long as the risk set is large in comparison to the number of events and the amount of censoring, the reason being that scenarios in which the allocation becomes highly unbalanced after some time (e.g. $Y_1\langle i\rangle = 80, Y_0\langle i\rangle = 20$) are extremely unlikely under the null.

Optimality close to hazard ratio 1 In case of balanced allocation, Figure 2.2 shows that the approximate e-values for a single event time from the Gaussian $M'_{\mu_1,0}\langle i\rangle$ are very similar to the exact e-values $M_{\theta_1,1}\langle i\rangle$ in designs for alternative hazard ratios θ_1 between

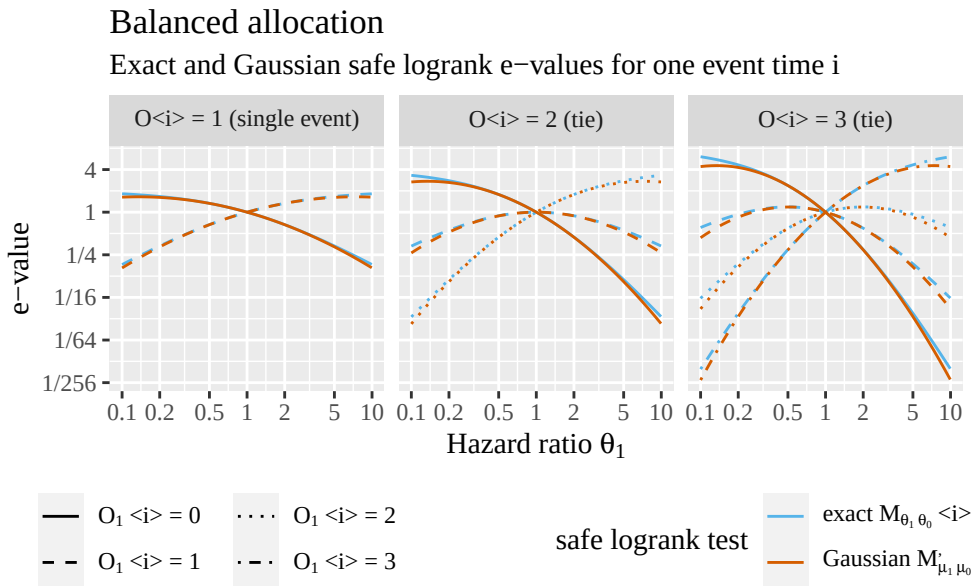

Figure 2.2. For balanced allocation ($m_1 = m_0 = Y_1\langle i\rangle = Y_0\langle i\rangle$) $M'_{\mu_1,\mu_0}\langle i\rangle$ is very similar to $M_{\theta_1,\theta_0}\langle i\rangle$ in case of designs for $0.5 \le \theta_1 \le 2$. Here $\theta_0 = 1$, $\mu_0 = 0$, and μ_1 follows from θ_1 as in (2.19). Note that both axes are logarithmic.

0.5 and 2. Having observed I event times, leading to a logrank statistic Z as in (2.16), we can therefore directly approximate $M^{\langle I\rangle}_{\theta_1,1}$ by

$$M''_{\mu_1,0} = \frac{\phi_{\mu_1\sqrt{n}}(Z)}{\phi_0(Z)} \quad \text{with } \mu_1 = \log(\theta_1)\cdot\sqrt{\frac{m_1\cdot m_0}{(m_1+m_0)^2}}. \tag{2.20}$$

This second approximation is valid under Schoenfeld's assumption (b) of constant large risk set. For then the initial risk set sizes m_0 and m_1 are both large in which case also the multiplicity correction in the definition of $V_1\langle i\rangle$ for ties is negligible. Without the multiplicity correction, the variance $V_1\langle i\rangle$ in (2.16) is equal to the variance of $O\langle i\rangle$ single events. Again, if the initial risk sets are sufficiently large, we can treat this variance to be constant in i and equal to $m_1 m_0/(m_1+m_0)^2$ per event. Taken together straightforward

calculus thus gives

$$\begin{aligned}\prod_{i=1}^{I} M'_{\mu_1,0}\langle i\rangle &= \exp\left(-\frac{1}{2}\sum_{i=1}^{I}\left(\mu_1^2 O\langle i\rangle - 2\mu_1\sqrt{O\langle i\rangle}Z\langle i\rangle\right)\right)\\ &\approx \exp\left(-\frac{1}{2}n\mu_1^2 + \mu_1\sum_{i=1}^{I}\sqrt{O\langle i\rangle}\,\frac{O_1\langle i\rangle - E_1\langle i\rangle}{\sqrt{O\langle i\rangle}\sqrt{\frac{m_1\cdot m_0}{(m_1+m_0)^2}}}\right)\\ &= \exp\left(-\frac{1}{2}n\mu_1^2 + \mu_1\frac{\sum_{i=1}^{I}\left\{O_1\langle i\rangle - E_1\langle i\rangle\right\}}{\sqrt{\frac{m_1\cdot m_0}{(m_1+m_0)^2}}}\right)\\ &\approx \exp\left(-\frac{1}{2}n\mu_1^2 + \mu_1\sqrt{n}Z\right) = M''_{\mu_1,0}.\end{aligned}$$

The first $\approx$ uses $\sum_{i=1}^{I} O\langle i\rangle = n$ and approximates the Z-score per event time from (2.18) assuming that the correction for ties is negligible and the risk set constant; the second $\approx$ maps back to the logrank statistic of multiple events Z using that the single event variance $m_1 \cdot m_0/(m_1+m_0)^2$ is approximately n times smaller than the variance $\sum_{i=1}^{I} V_1\langle i\rangle$ used to define Z in (2.16).

We will call $M''_{\mu_1,0}$ the *approximately safe logrank test based on the Gaussian approximation on the logrank statistic* or simply the *Gaussian safe logrank test*. In Figure 2.3 we investigate the power of this Gaussian safe logrank test and confirm that it often achieves e-value > 20 at exactly the same sample size as the exact safe logrank test. (In Section 2.3 we investigate this further.) Thus, a hazard ratio between 0.5 and 1 – and by symmetry also between 1 and 2 – is apparently sufficiently close to 1 (Schoenfeld's assumption (a)). Also, the deviations from the constant large and balanced risk set do not seem to occur often for this range of hazard ratios (Schoenfeld's assumption (b)). After all, the risk set needs to be quite large to observe the number of events to detect hazard ratios in the range $0.5 \leq \theta_1 \leq 2$. Figure 2.2 and Figure 2.3 show that the closer θ_1 is to 1, the more similar the Gaussian safe logrank test and the exact safe logrank test behave, but that the exact safe logrank test is optimal in general.

2.2 Comparing rejection regions

The first approaches to sequential analysis came with precise stopping rules. Wald's Sequential Probability Ratio Test (SPRT) (Wald, 1947), for example, specifies an upper and a lower boundary and requires the experiment to stop when either of the two is crossed. Hence the guarantees for type-I error control in crossing the upper boundary strictly rely on the lower boundary and vice versa.

Sequential analysis became popular in monitoring of clinical trials (by so-called Data Safety and Monitoring Boards, DSMBs) with the arrival of group-sequential methods – especially when timing interim analyses was flexibilized by α-spending functions (Lan

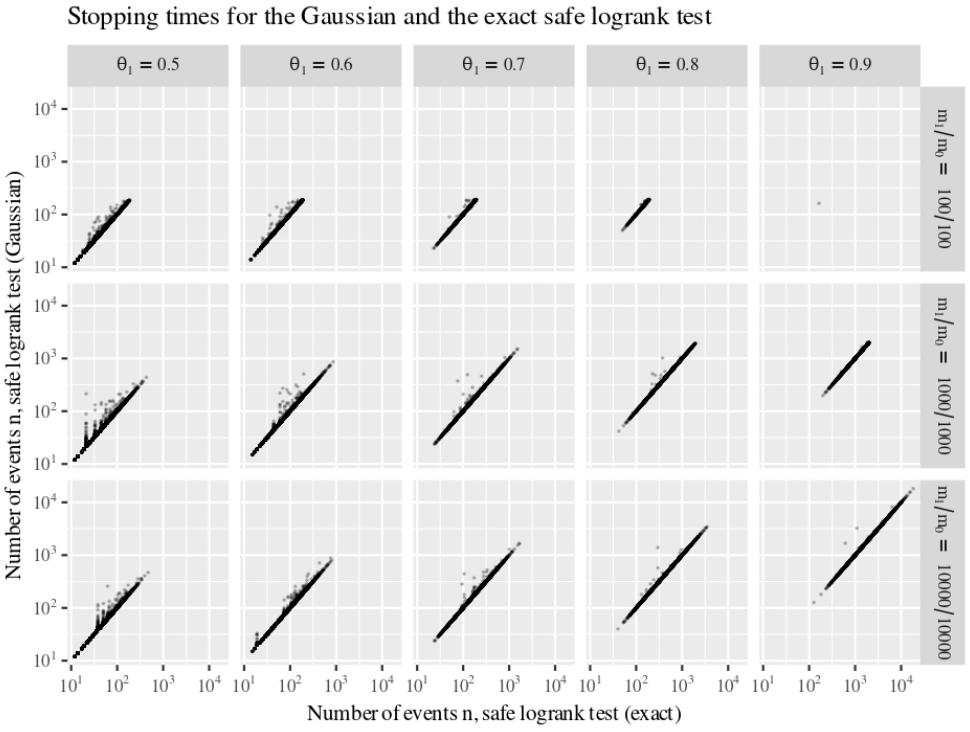

Figure 2.3. *Stopping times $I = n$ (number of events before stopping) simulated as single events to achieve e-value > 20 (safe test with $\alpha = 0.05$), when using the exact logrank martingale $M^{\langle n \rangle}_{\theta_1,1}$ from (2.10) and the Gaussian approximation on the logrank statistic $M''_{\mu_1,0}$ from (2.20). θ_1 specifies both the hazard ratio used for designing the tests and the true hazard ratio used to simulate the risk set process. Details of the simulation are given in Appendix Section 2.D. Note that both axes are logarithmic.*

and DeMets, 1983). Proschan et al. (2006, p. 214) note that "data are often not available in the order in which patients have been accrued", and therefore patient-by-patient or event-by-event analyses are not possible while monitoring the data. Moreover, boundaries are sometimes more guidelines than strict stopping rules: "One can regard a clinical trial that compares a new treatment to placebo or to an old treatment as having one clearly defined upper one-sided boundary – the one whose crossing demonstrates benefit – and a number of less well defined one-sided lower boundaries, the ones whose crossing worries the DSMB." (Proschan et al., 2006, p. 6). So two criteria are critical: (1) the analysis can separate benefit from harm, upper boundaries from lower boundaries – ignoring the crossing of one should not invalidate the inference from crossing the other; (2) crossing a boundary in the past should not impair a DSMB that can only convene for decisions at irregular intervals with chronologically incomplete data. These two criteria are fulfilled by α-spending functions, but in fact also by our safe logrank test: Ville's inequality (2.11)

allows for any stopping rule and while the test can be two-sided (2.14), it does not need to be.

Here we compare the region of logrank statistics for which α-spending approaches and the safe logrank test reject the null hypothesis of no effect (hazard ratio $\theta_0 = 1$). We discuss the two main α-spending functions that are inspired by two group-sequential approaches – Pocock (1977) and O'Brien and Fleming (1979) – although our main focus is on the O'Brien-Fleming approach. The Pocock procedure rejects at equally extreme values for the Z-statistic for small sample sizes as for larger ones, while DSMB's usually do not want to stop a trial very early for efficacy. Pocock himself now believes that his boundary is unsuitable (Pocock, 2006). Moreover, in contrast to the Pocock approach, the O'Brien Fleming α-spending approach can be used to monitor the data after each new event-time. Hence the fair assessment is to compare continuously monitoring the safe logrank test to continuously monitoring using the O'Brien-Fleming α-spending function.

2.2.1 GROW safe logrank (Gaussian) vs α-spending

In the previous section, Figure 2.2 and Figure 2.3 show that, in case of balanced allocation, the Gaussian approximation to the logrank statistic behaves very similar to the exact logrank test for certain designs. This is the case if we design a trial with 10000 participants to detect an effect of minimum clinical relevance of $\theta_1 = 0.7$ using a design that is *growth rate optimal in the worst case* (GROW, see Example 2). We will take this trial design as an example to compare the safe logrank test to α-spending. Since the O'Brien-Fleming rejection region can be uniquely defined in terms of the logrank Z-statistic, we compare it to the Gaussian safe logrank test defined on the same logrank statistic.

Gaussian safe logrank Z-rejection region For the Gaussian approximation we can specify the region of values for the logrank Z-statistic that rejects the null hypothesis when the e-value is larger than $1/\alpha$ as follows: rejection takes place for values of Z such that

$$M''_{\mu_1,\mu_0} = \frac{\phi_{\mu_1\sqrt{n}}(Z)}{\phi_{\mu_0}(Z)} = \frac{\exp[-\frac{1}{2}(Z-\mu_1\sqrt{n})^2]}{\exp[-\frac{1}{2}Z^2]} \geq \frac{1}{\alpha} \text{ with } \mu_0 = 0;\ \mu_1 = \log(\theta_1)\sqrt{\frac{m_1 \cdot m_0}{(m_1+m_0)^2}}$$

such that for $m_1 = m_0$, we reject if:
$$\begin{cases} Z \geq \frac{1}{2}\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n} - \frac{\log(\alpha)}{\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n}} & \text{if } \theta_1 > 1 \\ Z \leq \frac{1}{2}\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n} - \frac{\log(\alpha)}{\log(\theta_1)\cdot\frac{1}{2}\cdot\sqrt{n}} & \text{if } \theta_1 < 1. \end{cases} \tag{2.21}$$

O'Brien-Fleming α-spending Z-rejection region The O'Brien-Fleming α-spending function gives a boundary that is constant in $B(n/n_{\max}) = Z/\sqrt{n/n_{\max}}$, where $n_{\max}$ is a maximum number of events that has to be set in advance. Result 5.1 in Proschan et al. (2006)

gives that if we take $B(n/n_{\max})$ to be continuous Brownian motion:

$$P_{\mu_0}\Big[B(n/n_{\max}) > c \quad \text{for some } n \leq n_{\max}\Big] = 2P_{\mu_0}\Big[B(n_{\max}/n_{\max}) > c\Big] = 2P_{\mu_0}\Big[Z > c\Big] = \alpha$$

$$\text{such that} \quad \begin{cases} Z \geq \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n/n_{\max}}} & \text{in case of a right-sided test design} \\ Z \leq \frac{\Phi^{-1}(\alpha/2)}{\sqrt{n/n_{\max}}} & \text{in case of a left-sided test design.} \end{cases} \tag{2.22}$$

Figure 2.4 plots both the Gaussian safe logrank and the O'Brien-Fleming α-spending rejection regions. The two regions of 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. 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 2.4 shows the boundary of the Pocock spending function for 10 interims. Note that the definition of the safe 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, you also assume an effect size of minimal interest. A fixed sample size analysis designed to detect a minimum hazard ratio of 0.7 would need a number of events $n_{\max} = n = 195$ to achieve 80% power if the true hazard ratio is 0.7. A sequential analysis using α-spending needs a bit more, a maximum sample size of $n_{\max} = 205$ events for an O'Brien-Fleming spending function and $n_{\max} = 245$ for a Pocock spending function, when we design for 10 looks. We investigate the number of events needed by the safe logrank test in the next section (note that the test allows for unlimited monitoring, so there is no real equivalent to $n_{\max}$).

The benefit of a sequential approach is that if the data looks better than hazard ratio 0.7 we can detect that with a number of events that is smaller than this maximum sample size. Figure 2.5 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 safe logrank test is that the safe 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. While an α-spending approach designed to have 80% power will miss out on rejecting the null hypothesis in 20% of cases (the type-II error) – shown to stay green in Figure 2.5 – the safe logrank test can potentially reject all. In the sequences of 500 events in Figure 2.5, all but one sequence of Z-statistics could be rejected at a larger sample size by the safe logrank test. By increasing the sample size, the safe logrank test can 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 errors are controlled. Figure 2.5 shows two null sequences of Z-statistics with a true hazard ratio of 1 that are rejected by the O'Brien-Fleming α-spending region, but not by the safe logrank test. Here, the safe logrank test is a bit more conservative, since it is saving α for the future.

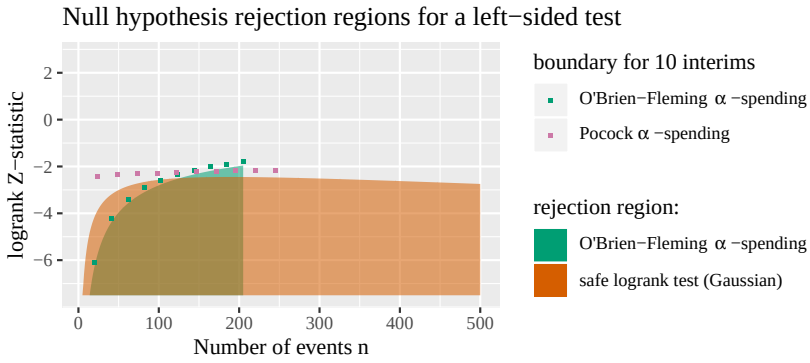

Figure 2.4. H_o rejection regions for continuously monitoring using O'Brien-Fleming α-spending and the safe logrank test, under balanced allocation ($m_0 = m_1$) and for one-sided $\alpha = 0.05$. Also shown are the O'Brien-Fleming and Pocock α-spending boundaries for 10 interim analyses. The α-spending boundaries are designed to have 80% power to detect a hazard ratio 0.7 (left-sided), leading to values of $n_{\max}$ given in the text. The safe logrank test is growth rate optimal for the worst case *hazard ratio $\theta_1 = \theta_{\min} = 0.7$ (left-sided), which we assume is of minimum clinical relevance.*

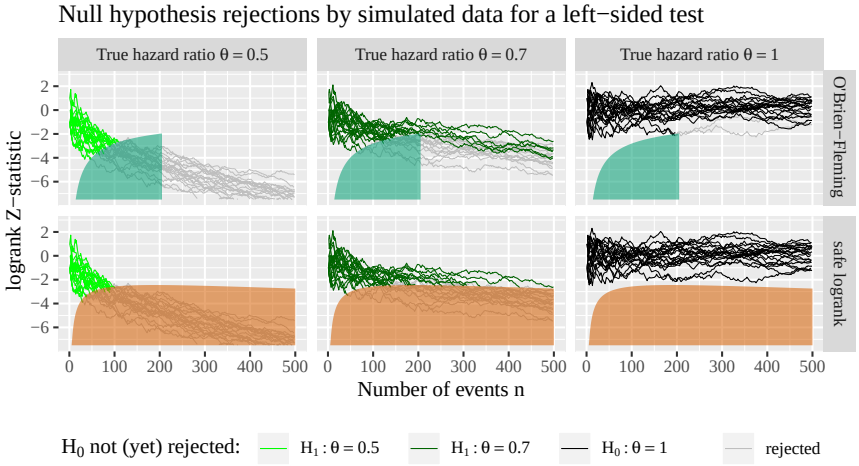

Figure 2.5. H_o rejected by simulated data in rejection regions as described in Figure 2.4 (designed to detect a hazard ratio of 0.7). The data is simulated under balanced allocation ($m_1 = m_0 = 5000$) and as time-to-event data such that ties can occur: the logrank Z-statistic does not have a value for all n; it sometimes jumps with several additional events at a time.

2.2.2 Unbalanced allocation

α-spending methods are known to behave poorly in case of unbalanced allocation (Wu and Xiong, 2017). In Section 2.1.3 we showed that our Gaussian approximation to the logrank test is also not an E-variable in case of unbalanced allocation. Our exact safe logrank test, however, is an E-variable under any allocation since it is defined directly on the risk set process (2.7), that takes into account the allocation. This suggests that in case of unbalanced allocation, the exact logrank test should be preferred over α-spending methods.

2.3 Comparing sample size

Figure 2.6 shows simulation results establishing three types of sample sizes. The leftmost panels ('Maximum') give the sample size required to design a trial. It expresses the maximum number of events $n_{\max}$ that needs to be observed under the alternative to achieve 80% power. In case of the classical logrank test and α-spending designs any further number of events beyond $n_{\max}$ cannot be analyzed. The rightmost panels ('Mean') show the sample size that captures the expected duration of the trial. It expresses the mean number of events, under the alternative, that will be observed before the trial can be stopped (here, for the safe logrank tests, we use the aggressive stopping rule that stops as soon as $M^{\langle n \rangle} \geq 1/\alpha = 20$ or $n = n_{\max}$). In case of α-spending approaches and the safe logrank test this number of events is always smaller than the maximum needed in the design stage. Finally, 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 – which will only happen 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 classical logrank test and O'Brien-Fleming α-spending are not recommended due to lack of type-I error control.

GROW safe logrank vs classical logrank and O'Brien-Fleming α-spending Figure 2.6 shows that the classical logrank test and O'Brien-Fleming α-spending require a smaller maximum number of events to obtain 80% power. This is the benefit of specifying a strict $n_{\max}$ in the design and spending all α such that you cannot analyze any further events. The additional number of events required for the maximum of the safe logrank test is the price to pay for the unlimited horizon, or for combining with the data from future trials, after some tial maximum has been reached (optional continuation). The Gaussian and exact safe tests are shown to be similar for $\theta_1 \geq 0.5$, but the Gaussian performs poorly for smaller hazard ratios. In terms of the mean number of events before the trial can be stopped, both the safe logrank test and O'Brien-Fleming outperform the classical standard logrank test at almost all hazard ratios for which the classical logrank test is recommended. The exact safe logrank test always needs more events than O'Brien-Fleming α-spending, but is the only approach that has exact type-I error control with small sample size at the smallest hazard ratios. The conditional mean number of events for the exact logrank test outperforms O'Brien-Fleming for designs with hazard ratio $\theta_1 = 0.1$

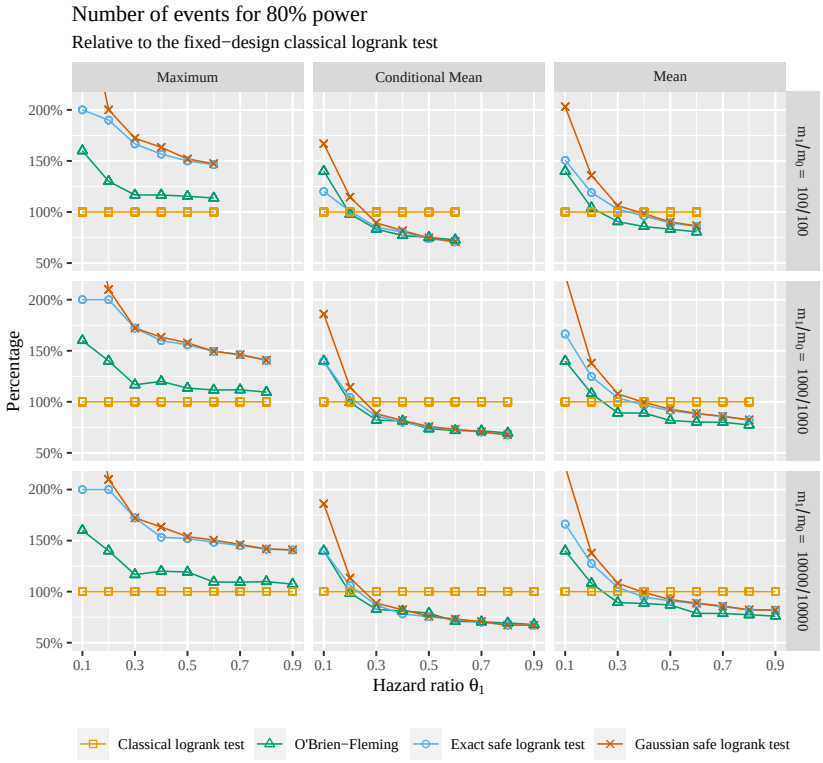

Figure 2.6. *Sample sizes (number of events) needed to reject the null hypothesis with* $\alpha = 0.05$ *using the GROW safe logrank test (exact or Gaussian,* $\theta_0 = 1$, $\mu_0 = 0$*), the classical logrank test (fixed sample size) and O'Brien-Fleming* α*-spending with continuous monitoring (see Section 2.2.1). All tests are designed to detect the hazard ratio* $\theta_1 = \theta_{\min}$ *on the x-axis and data is generated based on that same hazard ratio (see Example 2). 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. They are based on Schoenfeld's Gaussian approximation, which underestimates the number of events required for hazard ratios far away from 1 (e.g. 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. The difference between Maximum, Conditional Mean and Mean plots is explained in the main text, with further details in Appendix Section 2.D. The upshot is that at all hazard ratios at which the Gaussian approximation to the classical logrank test 'works' (say for* $\theta_1 \geq 0.3$*), the mean number of events needed by the safe logrank tests is about the same or noticeably smaller.*

and small risk set m_1/m_0 and behaves very similar under all other scenarios. While the mean of the stopping times (conditional mean) are similar, the rejection regions in Figure 2.4 illustrate, however, that the stopping times themselves are not the same. The safe logrank test will sometimes reject the null hypothesis at an earlier number of events than O'Brien-Fleming α-spending can, but also stop later than the $n_{\max}$ that was used to design the O'Brien-Fleming α-spending boundary.

2.4 Variations and extensions

2.4.1 Prequential Plug-In and Bayes predictive distributions for the alternative

Now suppose we do not have a very clear idea of which parameter $\theta_1 \in \Theta_1$ to pick. One way to handle this case is to use (2.12) rather than (2.8) and (2.10), replacing the fixed conditional probabilities $q_{\theta_1}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle)$ by a conditional probability mass function $r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle)$ that depends on the past and enables implicitly learning θ_1 from the data. For example, we may take

$$r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle) := q_{\hat{\theta}_i}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) \tag{2.23}$$

with q as in (2.6) and $\hat{\theta}_i := \hat{\theta}(\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle)$ the maximum likelihood estimate based on all past data, smoothed by adding two 'virtual' data points to the data, one for both groups:

$$\hat{\theta}(\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle) :=$$
$$\arg\max_{\theta\in(0,\infty)} \left(\prod_{k=1}^{i} q_\theta(O_1\langle k\rangle \mid Y_1\langle k\rangle, Y_0\langle k\rangle) \right) \cdot q_\theta(1 \mid Y_1\langle 1\rangle{+}1, Y_0\langle 1\rangle) \cdot q_\theta(0 \mid Y_1\langle 1\rangle, Y_0\langle 1\rangle{+}1).$$

Suppose the data are actually sampled from the process defined by q_θ, for some arbitrary but fixed θ. For sufficiently large initial risk sets (m_0, m_1 are not too small), by the law of large numbers, $\hat{\theta}_i$ will converge with high probability to θ, and $q_{\hat{\theta}_i}$ will behave more and more like the real q_θ from which data are sampled. Thus, the process (2.12) instantiated with (2.23) will behave more and more similarly to the 'correct' likelihood ratio (2.10). The process $r_1, r_2, \ldots$ is a typical instance of Dawid's (1984) *prequential plug-in* likelihood, that is often based on suitable smoothed likelihood-based estimators (Grünwald and Roos, 2020). Instead of r_i based on a plug-in estimate of θ based on $\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$, one may just as well use a Bayes predictive distribution based on the same data and some prior W_1 on θ. That is, we set

$$r_i(O_1\langle i\rangle \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle) := q_{W_i}(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) \tag{2.24}$$

where $W_i := W \mid \vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$ is the Bayes posterior on θ based on prior W and data $\vec{Y}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle$, and $q_W(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) := \int q_\theta(O_1\langle i\rangle \mid Y_1\langle i\rangle, Y_0\langle i\rangle) dW(\theta)$ is the Bayes predictive. By multiplying out the conditional probability mass functions r_i, we then get that $M^{\langle n\rangle}_{r,\theta_0} = \prod_{i=1}^{n} M_{r,\theta_0}\langle i\rangle$ is a Bayes factor between the Bayes marginal based on W and θ_0 (GHK explain how such a correspondence between Bayes factors and test martingales

holds more generally for simple null hypotheses). We do not know of a prior for which this Bayes factor or the constituent products have an analytic expression, but it can certainly be implemented using e.g. Gibbs sampling.

As explained underneath (2.12), the use of the r_i instead of q_{θ_1} does not compromise on safety: type-I errors and confidence sequences based on M_{r,θ_0} remain valid, whether the r_i are plug-in estimators or Bayes predictive distributions, no matter what prior W was chosen. Thus, our set-up is actually more intimately related to the concept of *luckiness* in the machine learning theory literature (Grünwald and Mehta, 2019) than to 'pure' Bayesian statistics. 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. Given a minimal clinically relevant effect size $\theta_{\min}$, the worst case logarithmic growth rate of M_{r,θ_0} will in general be less than that of the GROW $M_{\theta_{\min},\theta_0}$. Nevertheless, M_{r,θ_0} can come quite close to the optimal for a whole range of potentially data-generating θ and may thus sometimes be preferable over choosing $M_{\theta_{\min},\theta_0}$. More precisely, the use of a prior allows us to exploit favourable situations in which θ is even smaller (more extreme) than $\theta_{\min}$. In such situations, the GROW $M_{\theta_{\min},\theta_0}$ is effectively misspecified. By using r_i that learn from the data, we may actually get an E-variable that grows faster than the GROW $M_{\theta_{\min},\theta_0}$ which is fully committed to detecting the worst case $\theta_{\min}$.

In Figure 2.7 we illustrate such a situation where we start with 1000 participants in both groups. We generated data using different hazard ratios, and used a 'misspecified' $M_{\theta_1,1}$ that always used $\theta_1 = 0.8$. Note that while this is still the GROW (minimax optimal) martingale test under $H_1 = \{P_\theta : \theta_1 \leq 0.8\}$, if we knew the true θ, we could use the faster-growing test martingale $M_{\theta,1}$. We will call the test based on this latter martingale the *oracle* exact safe logrank test, since it is based on inaccessible (oracle) knowledge. We estimated the number of events that allows for 80% power for the tests based on $M_{0.8,1}$ and the oracle $M_{\theta,1}$ and the prequential plug-in $M_{r,1}$ with r as in (2.23). In all cases we used the aggressive stopping rule that stops as soon as $M_{\cdot,1} > 1/\alpha = 20$. We see that, as the true θ gets smaller than 0.8, we need less events using the GROW test $M_{0.8,1}$ (the data are favorable to us), but using the oracle exact safe logrank test we get a considerable additional reduction. The prequential plug-in $M_{r,1}$ 'tracks' the oracle $M_{\theta,1}$ by learning the true θ from the data: for θ near 0.8, it behaves worse (more data are needed) than $M_{0.8,1}$ (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 of Figure 2.6. Notably, the O'Brien-Fleming procedure, even though unsuitable for optional continuation, needs even more events than the misspecified safe logrank test $M_{0.8,1}$ as soon as θ goes below 0.8 (the simulations were performed using exactly the same algorithms as for Figure 2.6 so the y-axis at $\theta = 0.8$ coincides with that of Figure 2.6, but now with absolute rather than relative numbers); details are described in Appendix Section 2.D).

2.4.2 Anytime-valid confidence sequences

Standard tests give rise to confidence intervals by varying the null and 'inverting' the corresponding tests. In analogous fashion, test martingales can be used to derive *anytime-*

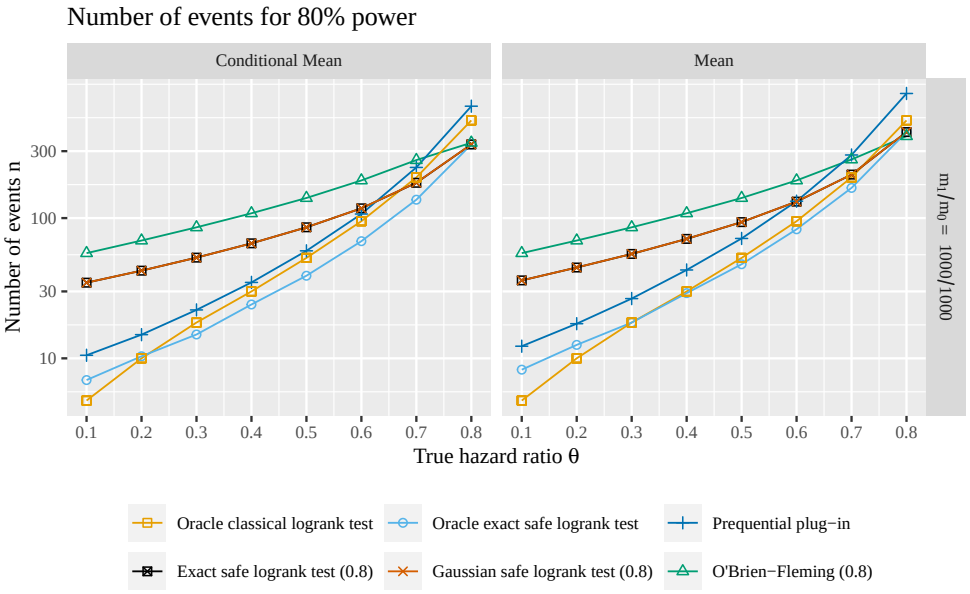

Figure 2.7. *We show the number of events at which one can stop retaining 80% power at* $\alpha = 0.05$ *using the process* M_{θ_1,θ_0} *with* $\theta_0 = 1$ *and* $\theta_1 = 0.80$ *when the true hazard ratio* θ *generating the data is different from* θ_1*. 'Oracle' means that the method is specified with knowledge of the true* θ*, which in reality is unknown. Note that the y-axis is logarithmic.*

valid (AV) confidence sequences (Darling and Robbins, 1967; Lai, 1976; Howard et al., 2018, 2021). In our setting, a $(1-\alpha)$-AV confidence sequence is a sequence of confidence intervals $\{\text{CI}_i\}_{i\in\mathbb{N}}$, one for each consecutive event, such that

$$P_\theta\,(\text{there is an } i \in \mathbb{N} \text{ with } \theta \notin \text{CI}_i) \leq \alpha. \tag{2.25}$$

A standard way to design $(1-\alpha)$- AV confidence sequences, translated to our logrank setting, is to use a prequential plug-in or Bayesian-based r as described in the previous subsection. After observing n events, one reports $\text{CI}_{n,\alpha} = [\theta_{n,L}, \theta_{n,U}]$ where $\theta_{i,L}$ is the largest θ_0 such that for all $\theta' \leq \theta_0$, $M^{\langle n\rangle}_{r,\theta'} \geq 1/\alpha$; similarly $\theta_{n,U}$ is the smallest θ_0 such that for all $\theta' \geq \theta_0$, $M^{\langle n\rangle}_{r,\theta'} \geq 1/\alpha$. That is, we check (2.12) where we vary θ_0 and we report the smallest interval such that $M_{r,\theta_0} > 1/\alpha$ outside this interval. Ville's inequality immediately shows that this gives an AV confidence sequence for arbitrary instantiations of r.

2.4.3 Covariates: the full Cox Proportional Hazards E-Variable

We extend the process of Section 2.1 (for now without ties) to explicitly represent participants, as done above Example 2 (and with the same notation as used there). Additionally, we now also fix a set of d covariates and let $\mathbf{Z}\langle i\rangle = (\vec{z}_1 \ldots \vec{z}_m)\langle i\rangle$ be the matrix consisting of the covariate vectors for each participant at the time of the i^{th} event:

$\vec{z}_j\langle i\rangle = (z_{j,1}\langle i\rangle, \dots, z_{j,d}\langle i\rangle)$. We let random variable $J\langle i\rangle$ denote the index of the patient to which the i^{th} event happens, and consider the extended process $J\langle 1\rangle, J\langle 2\rangle, \dots$ where the information that is available at time i is $\vec{g}, J\langle 1\rangle, \dots, J\langle i\rangle, \mathbf{Z}\langle 1\rangle, \dots, \mathbf{Z}\langle i\rangle$. In this section we re-define $\vec{Y}\langle i\rangle = (Y_1\langle i\rangle, \dots, Y_m\langle i\rangle) \in \{0,1\}^m$ to be a vector of m components, the j component indicating whether the j^{th} participant is still without an event just before the i^{th} event. Accordingly we set $\vec{Y}\langle 1\rangle = (1, \dots, 1)$ and we set $\vec{Y}_j\langle i+1\rangle = \vec{Y}_j\langle i\rangle$ for $j \neq J\langle i\rangle$, $\vec{Y}_j\langle i+1\rangle = 0$ for $j = J\langle i\rangle$. The conditional distribution underlying the process is now denoted $P_{\beta,\theta}$ with $\theta > 0$ and $\beta \in \mathbb{R}^d$, defined as follows: $P_{\beta,\theta}$ is given by, for $j \in \{1, \dots, m\}$ and $\vec{y} \in \{0,1\}^m$ with $\vec{y}_j = 1$:

$$P_{\beta,\theta}(J\langle i\rangle = j \mid \vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \dots, \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle) := P_{\beta,\theta}(J\langle i\rangle = j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)$$
$$P_{\beta,\theta}(J\langle i\rangle = j \mid \vec{Y}\langle i\rangle = \vec{y}, \mathbf{Z}\langle i\rangle = \mathbf{z}) := q_{\beta,\theta}(j \mid \vec{y}, \mathbf{z}) := \frac{\exp(\beta^T \vec{z}_j + \theta' g_j)}{\sum_{j':\vec{y}_{j'}=1} \exp(\beta^T \vec{z}_{j'} + \theta' g_{j'})}, \tag{2.26}$$

with $\theta' = \log\theta$ and $\mathbf{z} = (\vec{z}_1, \dots, \vec{z}_m)$. This is consistent with Cox' (1972) proportional hazards regression model: the probability that the j^{th} participant has an event, assuming he/she is still at risk, is proportional to the exponentiated weighted covariates, with group membership being one of the covariates. In case $\beta = 0$, this is easily seen to coincide with the definition of P_θ via (2.6).

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

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

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

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

We know from Li (1999) and GHK that $\tilde{q}_{\leftarrow W,0}(\cdot \mid \vec{y})$ exists. As explained by GHK in the context of E-variables for 2×2 contingency tables, the fact that the random variables $Y\langle i\rangle$ constituting our random process have finite range implies that, for each W, the infimum is in fact achieved by some distribution W' with finite support on $\mathbb{R}^d$. For given $\theta_0, \theta_1 > 0$, let

$$M_{W,\theta_1,\theta_0}\langle i\rangle = \frac{q_{W,\theta_1}(J\langle i\rangle \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)}{q_{\leftarrow W,\theta_0}(J\langle i\rangle \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)} \tag{2.27}$$

be our analogue of $M_{\theta_1,\theta_0}\langle i\rangle$ as in (2.8).

Theorem 2.4.1. [Corollary of Theorem 1 from GHK19] *For every prior W on $\mathbb{R}^d$, for all $\tilde{\beta} \in \mathbb{R}^d$,*

$$\mathbf{E}_{P_{\tilde{\beta},\theta_0}}[M_{W,\theta_1,\theta_0}\langle i\rangle \mid \vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle] = \sum_{j\in[m]} q_{\beta,\theta_0}(j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle) \cdot \frac{q_{W,\theta_1}(j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)}{q_{\leftarrow W,\theta_0}(j \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle)} \leq 1 \tag{2.28}$$

so that $M_{W,\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on $\vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \ldots, \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle$.

Note that the result does not require the prior W to be well-specified in any-way: under any $(\tilde{\beta}, \theta_0)$ in the null distribution, even if $\tilde{\beta}$ is completely disconnected to W, $M_{W,\theta_1,\theta_0}\langle i\rangle$ is an E-variable conditional on past data.

In particular, since the result holds for arbitrary priors, it holds, at the n^{th} event time, for the Bayesian posterior $W_{n+1} = W_1 \mid \vec{Y}\langle 1\rangle, \mathbf{Z}\langle 1\rangle, \ldots, \vec{Y}\langle n\rangle, \mathbf{Z}\langle n\rangle$, based on arbitrary prior W_1 with density w_1, i.e. the density of W_{n+1} is given by

$$w_{n+1}(\beta) \propto \prod_{i=1}^{n} q_{\beta,\theta}(J\langle i\rangle \mid \vec{Y}\langle i\rangle, \mathbf{Z}\langle i\rangle) w_1(\beta).$$

Using Definition Theorem 2.0.3, we can therefore, for each prior W_1, construct a test martingale $M^{\langle n\rangle} := \prod_{i=1}^{n} M_{W_i,\theta_1,\theta_0}$ that 'learns' β from the data, analogously to (2.24), and computes a new RIPr at each event time i.

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

Ties While in the case without covariates, our E-variables allowing for ties (Section 2.1.2) correspond to a likelihood ratio of noncentral hypergeometrics, the situation is not so simple if there are covariates—although deriving the appropriate extension of the non-central hypergeometric partial likelihood is possible, one ends up with a hard-to-calculate formula (Peto, 1972). Various approximations have been proposed in the literature (Cox, 1972; Efron, 1974). In case these preserve the E-variable and martingale properties, they would retain type-I error probabilities under optional stopping and we could use them without problems. We do not know whether this is the case however; for the time being,

we recommend handling ties by putting the events in a worst-case order, leading to the smallest values of the E-variable of interest, as this is bound to preserve the type-I error guarantees.

2.5 Discussion, Conclusion and Future Work

We introduced the safe logrank test, a version of the logrank test that can retain type-I error guarantees under optional stopping and continuation. Extensive simulations revealed that, if we do engage in optional stopping, it is competitive with the classical logrank test (which neither allows in-trial optional stopping nor optional continuation) and α-spending (which allows forms of optional stopping but not optional continuation). We provided an approximate test for applications in which only summary statistics are available and also showed how the safe logrank test can be used in combination with (informative) prior and prequential learning approaches, when no effect size of minimal clinical relevance can be specified. Two of our extensions invite further research: We introduced anytime-valid confidence sequences for the hazard ratio, and will study these more in future work into their performance in comparison to other approaches. We also introduced an extension to Cox' proportional hazards regression, which promises type-I error guarantees even if the alternative model is equipped with arbitrary priors. In future work, we plan to implement this extension – which requires the use of sophisticated methods for estimating mixture models. The GROW safe logrank tests (exact and Gaussian) are already available in our SafeStats R package (Turner et al., 2022). We end with two final points of discussion: *staggered entries* and *doomed trials*.

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 e.g. a two-day-after-surgery event changes when new participants enter and survive two days). Sellke and Siegmund (1983) and Slud (1984) show that, in general, martingale properties cannot be preserved under such staggered entry, but that asymptotic results are hopeful (Sellke and Siegmund, 1983) as long as certain scenarios are excluded (Slud, 1984). When all participants' risk is on the same (calendar) time scale (e.g. infection risk in a pandemic; 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 safe for patient time under staggered entry, and those that need extra care.

Your trial is not doomed In their summary of conditional power approaches in sequential analysis Proschan et al. (2006) 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" (Proschan et al., 2006, p. 63). While this is the

case for all existing sequential approaches that set a maximum sample size, this is not the case for safe 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 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.

Code availability

This chapter's R code is available on `https://osf.io/3n8g2/` (Ter Schure and Pérez-Ortiz, 2022).

Appendices

Appendix Section 2.A connects the simple risk set processes in this chapter to a continuous time survival process. Section 2.B gives an additional argument for the GROW criterion, Section 2.C gives a more detailed derivation of the logrank test as a score test for single events and ties and Section 2.D gives a step-by-step account of the sample size comparison simulations.

2.A Towards Continuous Time

In Section 2.1.1 and Section 2.4.3, the expression (2.26) and its simplified form (2.7) defining the safe logrank test appeared as a factor in a standard likelihood, defined relative to a basic stochastic process in which the time between events was not formalized. Cox (1975) simply claimed it to be also a partial likelihood for the underlying continuous-time process with proportional hazards determined by (β, θ'). Slud (1992) proved that it can indeed be seen as such, in the same general setting with time-varying continuous covariates that we consider here (see also Andersen et al. (1993) for closely related results). This already shows that the test martingales of Section 2.1.1 (no covariates, no ties) and Section 2.4.3 (covariates, no ties) also remain test martingales in the continuous time setting. The only part that is not covered by such existing results is the case of ties, Section 2.1.2: if we plug the conditional distributions (2.15) that allow for ties into the test martingale (2.8), do we still get a test martingale in continuous time if $\theta = 1$? Since we have not been able to find a complete proof in the literature of this result or a result that would directly imply it, we derive a simple version of such a result below, implicitly also proving the result without ties, though in a less general setting (without covariates, and assuming continuous hazards) than Slud. For simplicity we only treat the case without censoring again.

As in Section 2.4.3 and above Example 2, we identify each of the $m = m_0 + m_1$ participants by an index $j \in [m] := \{1, \dots, m\}$, with $\vec{g} \in \{0,1\}^m$, and $\vec{g}_j \in \{0,1\}$ denoting the group assignment of participant j. However, for simplicity we shall not make use of any covariates. For $t \in \mathbb{R}_0^+$ (continuous time), we let $T_j = t$ denote that the j^{th} participant had an event at time j; $T\langle i \rangle$ denotes the time at which the i-th event takes place. For $g \in \{0,1\}$, we let $Y_g(t) = \sum_{j:\vec{g}_j = g} Y_j(t)$ be the number of participants at risk in the group g at time t.

Fix some time $t^* > 0$. If patient with index j is in group g, then by assumption (2.13) we have, for fixed $0 < \epsilon < t^*$,

$$p_g^*(\epsilon) := P_g(t^* - \epsilon < T_j \leq t^* \mid T_j > t^* - \epsilon) = 1 - \exp\left(\int_{t^*-\epsilon}^{t^*} \lambda_g(s)\,\mathrm{d}s\right).$$

Let O_g^* be the number of patients of group $g \in \{0,1\}$ that witnessed the event of interest in the time interval $(t^* - \epsilon, t^*]$, and let $O^* = O_0^* + O_1^*$ be the total number of such patients.

Let $\mathscr{E}$ be the event denoting everything that is known just before time $t^*-\epsilon$, i.e. if k events happened before $t^*-\epsilon$, then $\mathscr{E}$ is the event that

$$T\langle 1\rangle = t_1, \ldots, T\langle k\rangle = t_k, Y_0\langle 1\rangle = y_0\langle 1\rangle, Y_1\langle 1\rangle = y_1\langle 1\rangle, \ldots, Y_0\langle k\rangle = y_0\langle k\rangle, Y_1\langle k\rangle = y_1\langle k\rangle$$

for specific $t_1, \ldots, t_k, y_0\langle 1\rangle, y_1\langle 1\rangle, \ldots, y_0\langle k\rangle, y_1\langle k\rangle$. Let $y_g := Y_g\langle k+1\rangle$ and note that y_g can be calculated at all times after $T\langle k\rangle$. By independence of the T_j, the distribution of O_g^* given $\mathscr{E}$ is binomial, given by, for ϵ, t^* such that $t^*-\epsilon > t_k$ and $o_g \leq y_g$,

$$P(O_0^* = o_0, O_1^* = o_1 \mid \mathscr{E}) = \\ \binom{y_0}{o_0} \cdot p_0^*(\epsilon)^{o_0}(1-p_0^*(\epsilon))^{y_0-o_0} \cdot \binom{y_1}{o_1} \cdot p_1^*(\epsilon)^{o_1}(1-p_1^*(\epsilon))^{y_1-o_1}.$$

Now let $\omega_g^*(\epsilon) = p_g^*(\epsilon)/(1-p_g^*(\epsilon))$. As is well known, the probability of (O_0^*, O_1^*) given $\mathscr{E}$ being binomial as above, the conditional probability of observing a particular O_1^* given $O^* = O_0^* + O_1^*$ and $\mathscr{E}$ must be given by Fisher's non-central hypergeometric distribution with parameter $\omega^*(\epsilon) = \omega_1^*(\epsilon)/\omega_0^*(\epsilon)$, whose probability mass function is given by, with ω abbreviating $\omega^*(\epsilon)$,

$$P(O_1^* = o_1 \mid \mathscr{E}, O^* = o) = q_\omega(o_1 | y_0, y_1, o) = \frac{\binom{y_1}{o_1}\binom{y_0}{o-o_1}\omega^{o_1}}{\sum_{\max\{0,o-y_1\}\leq u\leq \min\{y_1,o\}} \binom{y_1}{u}\binom{y_0}{o-u}\omega^u}.$$

We now note that if $\theta = 1$, then $\omega^*(\epsilon) = 1$ irrespective of ϵ and this reduces to the hypergeometric distribution with parameter $\theta = 1$ as in (2.15). This proves the claim made in Section 2.1.2: q_θ as in (2.15) with $\theta = 1$ is the correct conditional distribution under the continuous time model introduced above Example 2 with hazard ratio 1. It follows, using the reasoning in Example 1, that $M_{\theta_1,1}^{\langle I\rangle}$ as underneath (2.15) is a test martingale, and our test with $\theta_0 = 1$ is once again exact. Under hazard ratios $\theta \neq 1$, $\omega^*(\epsilon) \neq \theta$, so we do not have an exact supermartingale (and hence not an exact test) anymore. Still, as noted by Mehrotra and Roth (2001),

$$\lim_{\epsilon \downarrow 0} \omega^*(\epsilon) = \theta, \tag{2.A.1}$$

so, if we make observations at subsequent time points that are close enough to each other, all of our results still hold in an approximate sense.

From (2.A.1) we also see that

$$P(O_1\langle k+1\rangle = 1 \mid \mathscr{E}, T\langle k+1\rangle = t^*) = \lim_{\epsilon \downarrow 0} P(O_1^* = 1 \mid \mathscr{E}, O^* = 1) = q_\theta(o_1 \mid y_0, y_1),$$

with q_θ given as in (2.6). This shows that our original conditional distributions are correct as well in continuous time, now under each hazard ratio $\theta > 0$, not just $\theta = 1$, and $M_{r,\theta}$ gives an exact test martingale and hence an exact test for each θ, as long as there are no ties.

2.B Expected Stopping Time, GROW and Wald's Identity

Let P_{θ_0} represent our null model, and let, as before, the alternative model be given as $H_1 = \{P_{\theta_1} : \theta_1 \in \Theta_1\}$ with $\Theta_1 = \{\theta_1 : 0 < \theta_1 \leq \theta_{\min}\}$ for some $\theta_{\min} < 1$. Suppose we perform a level α test based on a test martingale M_{θ_1,θ_0} using the aggressive stopping rule: stop as soon as $M_{\theta_1,\theta_0} \geq 1/\alpha$. The GROW criterion (Chapter 2, Example 2) tells us to use $\theta_1 = \theta_{\min}$. Here we motivate this GROW criterion by showing that it minimizes, in a worst-case sense, the expected number of events needed before there is sufficient evidence to stop. The calculation below ignores the practical need to prepare for a bounded maximum number of events. For such more complicated considerations, we need to resort to simulations as in the main text.

We will further make the simplifying assumption that the initial risk sets (i.e. m_0 and m_1) are large enough so that for all sample sizes we will ever encounter, $Y_0\langle i\rangle / Y_1\langle i\rangle \approx m_0/m_1$. This allow us to act as if the random variables $O_1\langle i\rangle$ are i.i.d. Bernoulli: the ith event is sampled from $q_{\theta_1}\big(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)\big)$ for some $\theta_1 \in \Theta_1$, with q_{θ_1} as given by (2.6), which becomes independent of i if y_0/y_1 is replaced by m_0/m_1. Assuming i.i.d. data enables a standard argument based on Wald's (1947) identity, originally due to Breiman (1961). As said, we stop as soon as $M := M_{\theta_1,\theta_0} \geq 1/\alpha$ or when we run out of data, leading to a stopping time τ_{θ_1} (which we will denote by τ when θ_1 is clear from the context). Suppose first that we happen to know that the data comes from a specific $\theta \in \Theta_1$. Wald's identity now gives:

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

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

$$\frac{\mathbf{E}_{P'_\theta}[\log M^{\langle\tau\rangle}_{\theta_1,\theta_0}]}{\mathbf{E}_{P'_\theta}\left[\log \frac{p'_{\theta_1}(O_1\langle 1\rangle)}{p'_{\theta_0}(O_1\langle 1\rangle)}\right]} = \frac{\log\frac{1}{\alpha} + \text{VERY SMALL}}{\mathbf{E}_{P'_\theta}\left[\log \frac{p'_{\theta_1}(O_1\langle 1\rangle)}{p'_{\theta_0}(O_1\langle 1\rangle)}\right]} \tag{2.B.1}$$

with VERY SMALL between 0 and $\log|\theta_1/\theta_0|$, and $p'_{\theta_1}(O_1\langle 1\rangle) = q_{\theta_1}(O_1\langle 1\rangle \mid Y_1\langle 1\rangle, Y_0\langle 1\rangle)$. The first equality is just definition, the second follows because we reject *as soon as* $M^{\langle\tau\rangle}_{\theta_1,\theta_0} \geq 1/\alpha$, so $M^{\langle\tau\rangle}_{\theta_1,\theta_0}$ can't be smaller than $1/\alpha$, and it can't be larger by more than a factor equal to the maximum likelihood ratio at a single outcome (if we would not ignore the probability of stopping because we run out of data, there would be an additional small term in the numerator).

If we try to find the θ_1 which minimizes this, and – as is customary in sequential analysis – we approximate the minimum by ignoring the VERY SMALL part, we see that the expression is minimized by maximizing $\mathbf{E}_{P'_\theta}[\log \frac{p_{\theta_1}(Y\langle 1\rangle)}{p_{\theta_0}(Y\langle 1\rangle)}]$ over θ_1. The maximum is clearly achieved by $\theta_1 = \theta$; the expression in the denominator then becomes the KL divergence between two Bernoulli distributions. It follows that under θ, the expected number of outcomes

until rejection is minimized if we set $\theta_1 = \theta$. Thus, we use the GROW E-variable relative to $\{\theta\}$ as our actual E-variable. We still need to consider the case that, since the real H_1 is 'composite', as statisticians, we do not know the actual θ; we only know $0 < \theta \leq \theta_{\min}$. So we might want to take a worst-case approach and use the θ_1 achieving

$$\max_{\theta_1} \min_{\theta:0<\theta\leq\theta_{\min}} \mathbf{E}_{P'_\theta}\left[\log \frac{p'_{\theta_1}(O_1\langle 1\rangle)}{p'_{\theta_0}(O_1\langle 1\rangle)}\right],$$

since, repeating the reasoning leading to (2.B.1), this θ_1 should be close to achieving

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

But this just tells us to use the GROW E-variable relative to H_1, which is what we were arguing for.

2.C Logrank test as a score test

2.C.1 Logrank test statistic for single events and ties

Let $Y_1\langle i\rangle$ denote that number of participants in the risk set that are in the treatment group at the time of the i^{th} event, and analogously $Y_0\langle i\rangle$ for the number of participants at risk in the control group. Let $O_1\langle i\rangle$ and $O_0\langle i\rangle$ count the number of observed events in the treatment group and control group at the i^{th} event time. For single events $O_1\langle i\rangle = 1$ if the event occurred in the treatment group, and $O_1\langle i\rangle = 0$ if a single event occurred in the control group. We can extend this Bernoulli case to multiple simultaneous events (ties) in which case $O_1\langle i\rangle$ can be larger than 1. Here we discuss both cases (single event and ties) together, but we will discuss them separately later on. We define $Y\langle i\rangle = Y_1\langle i\rangle + Y_0\langle i\rangle$ and $O\langle i\rangle = O_1\langle i\rangle + O_0\langle i\rangle$.

Under the null hypothesis, at each event time i, the number of observed events in the treatment group $O_1\langle i\rangle$ follows a hypergeometric distribution with an expected number of events $E_1\langle i\rangle$ and a variance $V_1\langle i\rangle$ that depends on the risk set as follows:

$$A_1\langle i\rangle = \frac{Y_1\langle i\rangle}{Y\langle i\rangle}, \qquad E_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle, \qquad V_1\langle i\rangle = O\langle i\rangle \cdot A_1\langle i\rangle \cdot \big(1 - A_1\langle i\rangle\big) \cdot \frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}.$$

This is the formulation found in Cox (1972, equation (26)) with $\frac{Y\langle i\rangle - O\langle i\rangle}{Y\langle i\rangle - 1}$ a "multiplicity" correction or "correction for ties" (Klein and Moeschberger, 2006, p. 207). In case of single events with $O\langle i\rangle = 1$ the variance reduces to the Bernoulli variance:

$$V_1\langle i\rangle = A_1\langle i\rangle \cdot \big(1 - A_1\langle i\rangle\big) = E_1\langle i\rangle \cdot \big(1 - E_1\langle i\rangle\big).$$

As a test statistic, a logrank Z-statistic is constructed that is calculated either for the treatment or for the control group and has an approximate standard normal (Gaussian) distribution under the null hypothesis. This Z-statistic for the treatment group 1 is:

$$Z = \frac{\sum_i \{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_i V_1\langle i\rangle}}.$$

This test statistic is used to reject the null hypothesis at level α in favor of a one-sided alternative (e.g. $H_0 : \lambda_1(t\langle i\rangle) < \lambda_0(t\langle i\rangle)$ for some i) in case the value of the Z-statistic is smaller than z_α. You expect a negative value for the Z-statistic for a lower risk in the treatment group, so you need the α^{th} lower percentage point of the standard normal (Gaussian) distribution. A two-sided test can be constructed by comparing $|Z| \geq z_{\alpha/2}$.

2.C.2 Score test for the Bernoulli partial likelihood (single events)

In Section 2.1.1 we constructed a partial likelihood based on the probability mass function of a Bernoulli $y_1\theta/(y_0 + y_1\theta)$-distribution. Given an observed data set of participants at risk ($\vec{Y}\langle i\rangle$) at all event times i we can define a product of Bernoulli likelihoods. Following Cox (1972), we define these in terms of the logarithm of the hazard ratio, i.e. $\beta = \log\theta$ with $\theta = \lambda_1(t\langle i\rangle)/\lambda_1(t\langle i\rangle)$ for all event times $t\langle i\rangle$:

$$\mathfrak{L}\big(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle\big) = \prod_{i=1}^{n} q_\beta\big(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)\big),$$

where

$$q_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle)) = \left(\frac{Y_1\langle i\rangle \cdot \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)}\right)^{O_1\langle i\rangle} \left(\frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)}\right)^{1-O_1\langle i\rangle}.$$

This is the likelihood formulated for single events by Cox (1972) for his proportional hazards model in the two-sample case (a single categorical covariate indicating two groups).

Following the likelihood above, our loglikelihood is:

$$\begin{aligned} L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle) &= \sum_{i=1}^{n} \Big\{ O_1\langle i\rangle \cdot \Big(\log[Y_1\langle i\rangle] + \beta - \log\big[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\big]\Big) \\ &\qquad + \big(1 - O_1\langle i\rangle\big) \cdot \Big(\log[Y_0\langle i\rangle] - \log[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)]\Big)\Big\} \\ &= \sum_{i=1}^{n} \Big\{ O_1\langle i\rangle\beta + O_1\langle i\rangle \cdot \Big(\log[Y_1\langle i\rangle] - \log[Y_0\langle i\rangle]\Big) + \log[Y_0\langle i\rangle] \\ &\qquad - \log\big[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\big]\Big\}. \end{aligned}$$

If we omit the parts that do not depend on β, we get the expression in Cox (1972):

$$L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle) \propto \sum_{i=1}^{n} \Big\{ O_1\langle i\rangle\beta - \log\big[Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\big]\Big\}.$$

So if we take the score:

$$U(\beta) = \frac{\mathrm{d}L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta} = \sum_{i=1}^{n} \left\{ O_1\langle i\rangle - \frac{Y_1\langle i\rangle \cdot \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)} \right\},$$

with variance (by the observed Fisher information)

$$
\begin{aligned}
-U'(\beta) &= -\frac{\mathrm{d}^2 L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots, \vec{Y}\langle n\rangle)}{\mathrm{d}\beta^2} \\
&= \sum_{i=1}^{n} \left\{ \frac{\left(Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\right) \cdot Y_1\langle i\rangle \cdot \exp(\beta) - Y_1\langle i\rangle \cdot \exp(\beta) \cdot Y_1\langle i\rangle \cdot \exp(\beta)}{\left(Y_0\langle i\rangle + Y_1\langle i\rangle \cdot \exp(\beta)\right)^2} \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{Y_0\langle i\rangle Y_1\langle i\rangle \exp(\beta)}{(Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta))^2} \right\} = \sum_{i=1}^{n} \left\{ \frac{Y_1\langle i\rangle \exp(\beta)}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)} \frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle \exp(\beta)} \right\}
\end{aligned}
$$

such that, if we evaluate the score function at a hazard ratio θ of 1, so with $\beta = \exp(\theta) = \exp(1) = 0$, we get:

$$
\begin{aligned}
U(0) &= \sum_{i=1}^{n} \left\{ O_1\langle i\rangle - \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle} \right\} = \sum_{i=1}^{n} \left\{ O_1\langle i\rangle - \frac{Y_1\langle i\rangle}{Y\langle i\rangle} \right\} = \sum_{i=1}^{n} \{O_1\langle i\rangle - A_1\langle i\rangle\}, \\
-U'(0) &= \sum_{i=1}^{n} \frac{Y_1\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle} \cdot \frac{Y_0\langle i\rangle}{Y_0\langle i\rangle + Y_1\langle i\rangle} = \sum_{i=1}^{n} A_1\langle i\rangle \cdot (1 - A_1\langle i\rangle).
\end{aligned}
$$

$A_1\langle i\rangle = E_1\langle i\rangle / O\langle i\rangle = E_1\langle i\rangle$ in case of single events. Hence by standardizing the score function under the null hypothesis, we get the logrank statistic for single events:

$$
Z = \frac{\sum_{i=1}^{n} \{O_1\langle i\rangle - A_1\langle i\rangle\}}{\sqrt{\sum_{i=1}^{n} A_1\langle i\rangle \cdot (1 - A_1\langle i\rangle)}} = \frac{\sum_{i=1}^{n} \{O_1\langle i\rangle - E_1\langle i\rangle\}}{\sqrt{\sum_i V_1\langle i\rangle}}.
$$

2.C.3 Score test for the Fisher hypergeometric partial likelihood (tied events)

The Peto one-step estimator for odds ratios is build from a general score test for 2x2 tables. The appendix of Yusuf et al. (1985) gives description that we can use for the logrank score test in the case of ties.

If at the i^{th} event time $O\langle i\rangle$ events are observed, the expression in (2.15) describes the probability that $O_1\langle i\rangle$ of these events happen in the treatment group. The general description of the likelihood for the i^{th} event time that follows from this is:

$$
\begin{aligned}
s_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle) &:= \frac{p_0(O_1\langle i\rangle) \cdot \exp(O_1\langle i\rangle \cdot \beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k) \cdot \exp(k \cdot \beta)} \\
\text{with } O_1^{\min}\langle i\rangle = \max\{0, O\langle i\rangle - Y_0\langle i\rangle\}; \quad O_1^{\max}\langle i\rangle &= \min\{O\langle i\rangle, Y_1\langle i\rangle\},
\end{aligned}
$$

with $p_0(O_1\langle i\rangle)$ the null hypothesis probability of $O_1\langle i\rangle$ observed events in the treatment group. Our null hypothesis is a hypergeometric distribution:

$$
p_0(O_1\langle i\rangle) = P_0\left((O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle\right) = \frac{\binom{Y_1\langle i\rangle}{O_1\langle i\rangle} \cdot \binom{Y_0\langle i\rangle}{O\langle i\rangle - O_1\langle i\rangle}}{\binom{Y\langle i\rangle}{O\langle i\rangle}},
$$

such that $s_\beta = q_\beta$, our Fisher hypergeometric likelihood (2.15) in Section 2.1.2:

$$\begin{aligned} s_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle) &= \frac{p_0(O_1\langle i\rangle)\cdot \exp(O_1\langle i\rangle\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)} \\ &= \frac{\frac{\binom{Y_1\langle i\rangle}{O_1\langle i\rangle}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - O_1\langle i\rangle}}{\binom{Y\langle i\rangle}{O\langle i\rangle}}\cdot\exp(O_1\langle i\rangle\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle}\frac{\binom{Y_1\langle i\rangle}{k}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - k}}{\binom{Y\langle i\rangle}{O\langle i\rangle}}\cdot\exp(k\cdot\beta)} \\ &= \frac{\binom{Y_1\langle i\rangle}{O_1\langle i\rangle}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - O_1\langle i\rangle}\cdot\exp(O_1\langle i\rangle\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle}\binom{Y_1\langle i\rangle}{k}\cdot\binom{Y_0\langle i\rangle}{O\langle i\rangle - k}\cdot\exp(k\cdot\beta)} \\ &= q_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle). \end{aligned}$$

So our likelihood is:

$$\mathfrak{L}\left(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots\right) = \prod_i s_\beta(O_1\langle i\rangle \mid (Y_0\langle i\rangle, Y_1\langle i\rangle), O\langle i\rangle)$$

And the loglikelihood is:

$$L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots) = \sum_i \left\{ \log\left[p_0(O_1\langle i\rangle)\right] + O_1\langle i\rangle\cdot\beta - \log\left[\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)\right]\right\}.$$

And its score:

$$U(\beta) = \frac{\mathrm{d}L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots)}{\mathrm{d}\beta} = \sum_i \left\{ O_1\langle i\rangle - \frac{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot k\cdot\exp(k\cdot\beta)}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)}\right\},$$

with variance (by the observed Fisher information)

$$\begin{aligned} -U'(\beta) &= -\frac{\mathrm{d}^2 L(\beta \mid \vec{Y}\langle 1\rangle, \vec{Y}\langle 2\rangle, \ldots)}{\mathrm{d}\beta^2} \\ &= \sum_i \left\{ \frac{\sum_k p_0(k)\cdot\exp(k\cdot\beta)\cdot\sum_k p_0(k)\cdot k^2\cdot\exp(k\cdot\beta) - \left(\sum_k p_0(k)\cdot k\cdot\exp(k\cdot\beta)\right)^2}{\left(\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot\exp(k\cdot\beta)\right)^2}\right\}. \end{aligned}$$

If we evaluate the score function at a hazard ratio θ of 1, so with $\beta = \exp(\theta) = \exp(1) = 0$,

we get:

$$
\begin{aligned}
U(0) &= \sum_i \left\{ O_1\langle i\rangle - \frac{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot k}{\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)} \right\} = \sum_i \left\{ O_1\langle i\rangle - \sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\cdot k \right\} \\
&= \sum_i \left\{ O_1\langle i\rangle - O_1\langle i\rangle \cdot \frac{Y_1\langle i\rangle}{Y\langle i\rangle} \right\} = \sum_i \{O_1\langle i\rangle - E_1\langle i\rangle\} \\
U'(O) &= \sum_i \left\{ \frac{\sum_k p_0(k)\cdot\sum_k p_0(k)\cdot k^2 - \left(\sum_k p_0(k)\cdot k\right)^2}{\left(\sum_{k=O_1^{\min}\langle i\rangle}^{O_1^{\max}\langle i\rangle} p_0(k)\right)^2} \right\} \\
&= \sum_i \left\{ \sum_k p_0(k)\cdot k^2 - \left(\sum_k p_0(k)\cdot k\right)^2 \right\} = \sum_i V_1\langle i\rangle.
\end{aligned}
$$

2.D 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 2.6 and Figure 2.7 in Section 2.3 and Section 2.4. 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 safe logrank, its Gaussian approximation, and its prequential-plugin variation). Finally, we explain how the numbers 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 risk set process from Section 2.1.1. Indeed, if we are testing some fixed θ_1 with $\theta_1 \leq 1$ against $\theta_0 = 1$, the odds of next event at the i^{th} event time happening in group 1 are $\theta_1 Y_1\langle i\rangle : Y_0\langle i\rangle$ under the alternative hypothesis, which is the hypothesis we sample from. Thus, simulating in which group the next event happens only takes a (biased) coin flip. We consider the one-sided testing scenario θ_1 (for some $\theta_1 \in (0,1)$) vs. $\theta_0 = 1$, and we fix our desired level to $\alpha = 0.05$. For each test martingale $M_{\theta_1,1}^{\langle n\rangle}$ of interest we first consider the stopping rule $\tau = \inf\{n : M_{\theta_1,1}^{\langle n\rangle} \geq 1/\alpha\}$, that is, we stop as soon as $M_{\theta_1,1}^{\langle n\rangle}$ crosses the threshold $1/\alpha$.

In order 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 τ'. Under τ' we stop at the first of two moments: either when our test martingale $M_{\theta_1,1}^{\langle n\rangle}$ 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 τ' given by $\tau' = \min(\tau, n_{\max})$. In those cases in which the test based on the stopping rule τ achieves a power higher than $1-\beta$ (for a type-II error rate β), 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 τ'. 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,

$$P_{\theta_1}(\tau \geq n) \leq 1-\beta.$$

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

In order to estimate $n_{\max}$ for a given risk-set sizes m_1, m_0 and alternative hypothesis hazard ratio θ_1, we sampled 10^4 realizations of the survival process (under θ_1) 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. We then computed the $(1-\beta)$-quantile of the observed empirical distribution of τ, and reported it as an estimate of the number of events $n_{\max}$ in the 'maximum' column in Figure 2.6.

We assessed the uncertainty in the estimation $n_{\max}$ using the bootstrap. We performed 1000 bootstrap rounds on the sampled empirical distribution of τ, 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 2.6 and Figure 2.7.

In the 'mean' column of Figure 2.6 and Figure 2.7 we plotted an estimate of the expected number of events $\tau' = \min(\tau, n_{\max})$. 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 plotted an estimate of $\tau' \mid \tau' < n_{\max}$, 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 (1981), 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 classical, and the number of events required is $n^S_{\max} = 4(z_\alpha + z_\beta)^2 / \log^2 \theta_1$, 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 $n^{OF}_{\max}$ needed to achieve 80% as follows. For each experimental setting (m_0, m_1, θ), we generated 10^4 realizations of the survival process under the alternative hypothesis under consideration and computed the corresponding trajectories of the logrank statistic. For each possible value n of $n^{OF}_{\max}$, we computed the fraction of trajectories for which O'Brien-Fleming's procedure correctly stopped when used with the maximum number of events set to n. We report as an estimate of the true $n^{OF}_{\max}$ the first value of n for which this fraction is higher than 80%, our predefined power.