On information fraction for Fleming-Harrington type weighted log-rank tests in a group-sequential clinical trial design

Madan G. Kundu[1] and Jyotirmoy Sarkar[2]

[1]Daiichi-Sankyo Inc., Basking Ridge, New Jersey, USA

[2]Indiana University-Purdue University Indianapolis, Indiana, USA

Abstract

When comparing survival times of treatment and control groups under a more realistic non-proportional hazards scenario, the standard log-rank (SLR) test may be replaced by a more efficient weighted log-rank (WLR) test, such as the Fleming-Harrington (FH) test. Designing a group-sequential clinical trial with one or more interim looks during which a FH test will be performed, necessitates correctly quantifying the information fraction (IF). For SLR test, IF is defined simply as the ratio of interim to final numbers of events; but for FH test, it can deviate substantially from this ratio. In this paper, we separate the effect of weight function (of FH test) alone on IF from the effect of censoring. We have shown that, without considering the effect of censoring, IF can be derived analytically for FH test using information available at the design stage and the additional effect due to censoring is relatively smaller. This paper intends to serve two major purposes: first, to emphasize and rationalize the deviation of IF in weighted log-rank test from that of SLR test which is often overlooked (Jiménez, Stalbovskaya and Jones); second, although it is impossible to predict IF for a weighted log-rank test at the design stage, our decomposition of effects on IF provides a reasonable and practically feasible range of IF to work with. We illustrate our approach with an example and provide simulation results to evaluate operating characteristics.

Keywords: Censoring distribution against events (CDE), delayed effects, early separation, Fleming-Harrington (FH) test, information fraction, interim analyses, late separation, non-proportional hazards, time-to-event endpoint, type-I error.

1 Introduction

In many comparative time-to-event (TTE) analyses, an implicit assumption is that the survival functions exhibit proportional hazards (PH). For example, PH assumption is essential to perform a standard log-rank (SLR) test or to fit a Cox proportional hazards regression model. However, in recent years, multiple clinical trials (e.g., Glioblastoma [1], acute lymphoblastic leukemia [2], head and neck cancer [3], and many others [4]) have exhibited late separation of survival curves, possibly due to a delayed treatment effect. Alternatively, survival curves may converge after early separation in some clinical trials (e.g., chemotherapy [5] and surgical and medical therapy [6]). Under these non-PH circumstances, to compare the survival distributions, the SLR test still remains a valid option, but it does not necessarily remain the most powerful test in the class of all linear rank tests. Moreover, ignoring the non-PH nature of the survival distributions may have serious consequences

This is the author's manuscript of the article published in final edited form as:

Kundu, M. G., & Sarkar, J. (2021). On information fraction for Fleming-Harrington type weighted log-rank tests in a group-sequential clinical trial design. Statistics in Medicine, 40(10), 2321–2338. https://doi.org/10.1002/sim.8905

on the outcomes of the clinical trial, including failure to meet statistical significance [7–9].

Under a non-PH scenario (i.e., when hazards ratio (HR) changes over time), to increase power, the SLR test should be replaced by the weighted log-rank (WLR) test, which puts non-uniform weights to different events [10–12]. One choice of weights proposed by Fleming and Harrington (FH) [13, 14], and denoted by $FH(\rho, \gamma)$ (see Section 2.2), has become very popular because of its flexibility to adapt to various non-PH scenario between by appropriately tuning ρ and γ. For example, $FH(\rho = 0, \gamma > 0)$ put greater weights to later events whereas, $FH(\rho > 0, \gamma = 0)$ assigns greater weights to earlier events. Of course, $FH(\rho = 0, \gamma = 0) \equiv 1$ refers to equal weights given to all events, which corresponds to the SLR test. More importantly, the FH test (that is, the WLR test with FH weights) has the following two desirable statistical properties: (1) It maintains the probability of type-I error when ρ and γ are pre-specified (Tsiatis [15]); (2) Empirical studies [7, 8] have shown that while the FH test suffers only a minimal loss of power under the PH scenario, it achieves a substantial gain in power under the non-PH scenario. Despite these advantages, the WLR test must be used with caution as it might lose power when weights are mis-specified [16].

To detect early success or futility in a clinical trial, often a group-sequential design is adopted in which one or more interim analyses are permitted before the final analysis. Validity of a group-sequential design is primarily based on two principles [17]: (a) Re-parameterization of the test-statistic as a Brownian motion with independent increments; (b) Determination of the group-sequential boundaries based on information fraction (IF) accrued at each interim look. Tsiatis [15] has shown that the members of the FH class of statistics can be re-parameterized as Brownian motions for group-sequential monitoring. However, estimation of IF, which determines the rejection boundaries for testing at the interim analysis, remains a challenging problem: For the SLR test, the IF is proportional to the number of interim events; but in general, this relationship is not guaranteed to hold for the WLR test [18]. As Brummel and Gillen [19] pointed out, if IFs accrued at the interim analyses are not correctly estimated at the design stage, then power and type-I error will differ from the originally targeted value: If IF is under-estimated, then the overall size of the test may be inflated; on the contrary, if IF is over-estimated, then the overall power of the test may be compromised. The focus of this paper is to correctly estimate IF for $FH(\rho, \gamma)$ test in a group-sequential trial.

Previous works [18–20] observed that for the $FH(\rho, \gamma)$ test, information growth depends on the underlying pooled survival and censoring distributions across treatment arms along with the number of events at interim and final analyses. Consequently, all previous works on estimating IF are primarily based on parametric assumptions on TTE, enrollment, and lost-to-follow-up (LTFU) distributions [18–21]. In practice, these distributions cannot be predicted upfront at the design stage. Jiménez, Stalbovskaya and Jones [22] proposed an approach to implement FH test in a group-sequential design which does not allow rejection of null hypothesis at the interim. Moreover, this approach did not account for the deviation in IF from the ratio of events, and as such compromised the overall size of the test when weighted for late separation [23]. In this paper, we recognize that IF cannot be predicted precisely upfront at the design stage due to uncertainty in enrollment, TTE and LTFU distributions. Therefore, we focus on obtaining a reasonable and practically feasible range of IF to work with. Our approach has two advantages: first, the uncertainty in distributions are reflected through the range of IF; second, the minimum IF can be used to define rejection boundaries at interim analyses without inflating the overall type I error.

In this paper, we separate the effect of the weight function $FH(\rho, \gamma)$ on IF from the effect of censoring combinedly introduced by enrollment, LTFU and TTE distributions. The impact of non-uniform weighting of events on IF is easy to understand: For example, if a weight function gives larger weights to later events to account for late separation, the accrual of information will be relatively slower at the beginning causing IF to be smaller than the ratio of interim to final number of events. Similarly, one can visualize the effect of a weight function that puts higher weights to earlier events or to events in the middle. In the first step, we measure the effect of the given weight function $FH(\rho, \gamma)$ alone on IF ignoring the effect of censoring (see Section 3.1). Aside from explaining the effect of a weight function alone on IF, this approach has two advantages: (1) Without considering the effect of censoring, IF can be derived analytically; (2) It represents maximum IF and minimum IF under late separation and early separation, respectively. In the second step, we estimate the effect of censoring realized through enrollment, LTFU and TTE distributions (see Section 3.2). The intuition for the effect of censoring on IF is as follows: FH weights are function of the survival function $S(\cdot)$; and censoring causes a larger drop in $S(\cdot)$ following an event. In this context, we define "censoring distribution against events" (CDE) as the distribution of censored patients observed between two successive events. The advantage of using CDE is that it reduces the three dimensional distributional problem of enrollment, TTE and LTFU into a single dimension. As it turns out, the additional effect of censoring on IF is relatively smaller than the effect of variable weighting alone, which in turn helps to construct reasonable and practically feasible range for IF.

In this paper, we focus on estimating IF for the FH test with emphasis on late separation; however, the same principle applies to any WLR test and the findings extend to early separation. In Section 2, we describe the clinical trial setting, and give an overview of the FH test and group-sequential monitoring. Determination of IF due to variable weighting without censoring is described in Section 3.1. The effect of censoring on IF, along with CDE, is presented in Section 3.2. The proposed approach is illustrated in Section 4 and simulation results are presented in Section 5. The overall implementation strategy of FH test in a group-sequential clinical trial design is presented in Section 6. The `R` codes to calculate the IF for FH tests and CDE are presented in the Appendix.

2 Preliminaries

2.1 Group-sequential setting

Consider a typical two-armed, randomized, group-sequential, clinical trial that compares a treatment arm with a control arm using time-to-event as a primary variable of interest. Altogether N patients are enrolled and randomized with $a : 1$ (treatment to control) allocation ratio in the study during the time interval $[0, E]$, measured in months, say. For simplicity, we assume equal allocation ratio between the arms. The data are analyzed not just at the planned end of the study, but also at $M \geq 1$ earlier time-points during the course of the study. At each of the $M + 1$ analysis time-points, the following hypotheses are tested:

$$H_0 : \theta(t) = 1 \text{ vs. } H_1 : \theta(t) < 1 \qquad \text{for all } t$$

where $\theta(t)$ denotes the treatment-to-control HR at time t after randomization. When $\theta(t) \equiv \theta$ (i.e., HR is constant over time), SLR test would be most powerful. However, SLR test may not be the most powerful when HR varies over time. The m^{th} $(m = 1, \cdots, M)$ interim analysis is carried out

after observing D_{IA_m} events. After each interim analysis, a decision is made either to stop the study due to early efficacy; or to continue the trial on to the next analysis. The final analysis is conducted after observing D_{FA} events. In principle, we have $D_{IA_1} < \cdots < D_{IA_M} < D_{FA}$. The total study duration is L (i.e., time between enrollment of first patient and observation of D_{FA} events). Further, assume that the events in the study are observed at follow-up times $t_1 < t_2 < \cdots < t_K$, with n_k patients at risk and d_k events observed at time t_k $(k = 1, \cdots, K)$.

2.2 A brief review of Fleming-Harrington (FH) test

Let's assume that at the time of analysis, the distributions of observed events on the two arms are as follows: At time t_k, d_{0k} events were observed from n_{0k} patients at risk on the control arm, and d_{1k} events were observed from n_{1k} patients at risk on the treatment arm. Let $d_{0k} + d_{1k} = d_k$ and $n_{0k} + n_{1k} = n_k$. Then any WLR test statistic can be expressed as (e.g., see [21])

$$\text{WLR test statistic} = \frac{S_{WL}}{\sqrt{Var(S_{WL})}},$$

$$\text{where} \quad S_{WL} = \sum_k W(t_k)[d_{1k} - E(d_{1k})] \qquad \text{with} \quad E(d_{1k}) = \frac{n_{1k} d_k}{n_k}$$

$$\text{and} \quad Var(S_{WL}) = \sum_k W^2(t_k) Var(d_{1k}), \quad \text{with} \quad Var(d_{1k}) = \frac{n_{0k} n_{1k} d_k (n_k - d_k)}{n_k^2 (n_k - 1)}$$

For an FH test with weight $FH(\rho, \gamma)$, an event observed at time t is given the weight

$$W(t) = [S(t)]^\rho \, [1 - S(t)]^\gamma, \quad \text{with} \quad \rho, \gamma \geq 0 \tag{1}$$

where $S(t)$ is the estimated survival probability at time t in the pooled population (both arms combined). Clearly, when $\rho = \gamma = 0$, the FH test reduces to the SLR test with weight $W(t) \equiv 1$. To account for late separation, $FH(\rho = 0, \gamma > 0)$ is used to put greater weights on the later events. Similarly, $FH(\rho > 0, \gamma = 0)$ is used to put greater weights on the earlier events. As noted earlier, the FH test maintains type I error probability when ρ and γ are pre-specified and can achieve substantial gain in power under a non-PH scenario.

2.3 Group-sequential monitoring

The key to formulating a group-sequential testing in a clinical trial design is to approximate the test statistic over time by the partial sum of independent and identically distributed random variables (that is, S-processes)[17]. Tsiatis [15] has shown that such an S-process can be constructed in the context of FH test which can be defined at the k^{th} observed event time-point as

$$S_{WL,k} = \sum_{r=1}^{k} W(t_r)[d_{1r} - E(d_{1r})]$$

where $W(t)$ is defined in (1). The next step is to define a Brownian motion process with independent increments, as in [24], as

$$B_k = \frac{S_{WL,k}}{\sqrt{I_k}}, \quad \text{with} \quad I_k = Var(S_{WL,k})$$

where I_k is the information accrued upto k^{th} observed event; and it equals $Var(S_{WL,k})$. Accordingly, information fraction (IF) after observing the k^{th} event is expressed as

$$IF(k) = \frac{I_k}{I_K}$$

where I_K is the maximum information available after observing all D_{FA} events in the study. Of the many approaches to compute the group-sequential boundaries based on IF accrued at each analysis timepoint that preserve the overall type-I error probability, the error spending function of Lan and DeMets [25] is the most popular.

3 Information fraction for a Fleming-Harrington test

As noted in Section (2.3), the pivotal component in group-sequential monitoring is to estimate the IF at each analysis time-point. Under the assumptions [21] of (a) no ties (i.e., $d_k = 1$), and (b) the ratio of individuals at risk on the two treatment arms at each observed event time is close to the original allocation ratio (i.e., $n_{0k}/n_{1k} \approx 1/a$), the expression of $Var(d_{1k})$ simplifies to

$$Var(d_{1k}) = \frac{n_{0k}}{n_k} \cdot \frac{n_{1k}}{n_k} \approx \frac{1}{a+1} \cdot \frac{a}{a+1} = \frac{a}{(a+1)^2}$$

and, therefore, the expression of $S_{WL,k}$ simplifies to

$$I_k = Var(S_{WL,k}) \approx \frac{a}{(a+1)^2} \sum_{r=1}^{k} W^2(t_r)$$

Thus, the information contributed by the k^{th} event, after substituting FH weights in Eq. (1), is

$$\text{Information contributed by } k^{th} \text{ event: } \frac{a}{(a+1)^2} W^2(t_k) = \frac{a}{(a+1)^2} [S(t_k)]^{2\rho}[1 - S(t_k)]^{2\gamma} \quad (2)$$

Consequently, given D_{IA} and D_{FA} events at interim and final analysis respectively, the IF becomes

$$IF(D_{IA}, \rho, \gamma) = \frac{\sum_{k=1}^{D_{IA}} [S(t_k)]^{2\rho}[1 - S(t_k)]^{2\gamma}}{\sum_{k=1}^{D_{FA}} [S(t_k)]^{2\rho}[1 - S(t_k)]^{2\gamma}} \quad (3)$$

For the SLR test (i.e., $\rho = 0, \gamma = 0 \implies W(t) \equiv 1$), the IF in Eq. (3) reduces to D_{IA}/D_{FA}. For the FH test, the IF can be very different from D_{IA}/D_{FA} as the observed events do not contribute to the information equitably. For example, if we weight the later events more heavily than the earlier events to account for late separation, information accrues at a slower rate compared to the equal weight scenario resulting in $IF < D_{IA}/D_{FA}$. Similarly, when we account for early separation, $IF > D_{IA}/D_{FA}$.

One way to estimate the IF is by plugging in the KM estimate of $S(\cdot)$ in Eq. (3). Let t_k denote the observed time of the k^{th} event and n_k^c denote the number of patients censored in $(t_{k-1}, t_k]$. Then the KM estimate of $S(\cdot)$ at time t_k is

$$\hat{S}(t_k) = \hat{S}(t_{k-1}) \left(1 - \frac{1}{N - k - n_k^c + 1}\right) = \prod_{r=1}^{k} \left(1 - \frac{1}{N - r - n_r^c + 1}\right) \quad (4)$$

Since censoring affects $S(\cdot)$, it should also affect expression of IF in Eq. (3). Therefore, IF for FH test, is impacted by two factors: (a) non-uniform weighting of events, and (b) censoring of patients. We separate the effect of these two factors and are described in Section 3.1 and Section 3.2, respectively.

3.1 Information fraction under variable weighting, without the effect of censoring

To evaluate the sole effect of the weight function on IF, we assume that minimum censoring time is greater or equal to maximum follow-up time at final analysis (say, L_{FA}) in the study implying $n_k^c = 0, \forall k = 1, \cdots, D_{FA}$. Under this assumption,

$$\hat{S}(t_k) = \hat{S}(t_{k-1})\left(1 - \frac{1}{N-k+1}\right) = \prod_{r=1}^{k}\left(1 - \frac{1}{N-r+1}\right) = \frac{N-k}{N} \tag{5}$$

Substituting Eq. (5) into Eq. (3), we obtain the IF with FH(ρ, γ) without censoring as

$$IF_0(D_{IA}, \rho, \gamma) = \frac{\sum_{r=1}^{D_{IA}} (N-r)^{2\rho} r^{2\gamma}}{\sum_{r=1}^{D_{FA}} (N-r)^{2\rho} r^{2\gamma}} \tag{6}$$

Clearly, the IF without censoring can be derived analytically based on N, D_{IA} and D_{FA}. Furthermore, when accounting for late separation [i.e., $\rho = 0$], the IF depends only on D_{IA} and D_{FA}, but not on N, since

$$IF_0(D_{IA}, \rho = 0, \gamma) = \frac{\sum_{r=1}^{D_{IA}} r^{2\gamma}}{\sum_{r=1}^{D_{FA}} r^{2\gamma}} \tag{7}$$

The `R` code to calculate IF without considering the effect of censoring is presented in Appendix 1.

3.2 Information fraction with censoring

3.2.1 Effect of censoring on information fraction

Patients might experience potentially two types of censoring—administrative censoring due to staggered enrollment, and LTFU censoring due to patients dropping out. If the final analysis time is L, patients who enrolled at the start can have follow-up time upto L, whereas patients who enrolled at time E can only have a shorter follow-up time up to $(L-E)$. Therefore, no patients would be administratively censored before a follow-up time of $L-E$. Let l_0 be the earliest follow-up time when censoring starts affecting $\hat{S}(t)$, as shown in Figure 1. This l_0 can be interpreted as smallest censoring time as well. In the absence of LTFU censoring, $l_0 = L - E$. However in presence of LTFU, l_0 can be much smaller than $L - E$.

The estimated survival function $\hat{S}(t)$ is affected by censoring because it reduces the pooled risk set, with heavier censoring resulting in a larger drop in $\hat{S}(\cdot)$ at each observed event time. If n_k^c patients are censored during time interval $(t_{k-1}, t_k]$ and N_k patients are at risk after the $(k-1)^{st}$ event, then

$$\hat{S}(t_k | n_k^c > 0) = \hat{S}(t_{k-1}) \quad 1 - \frac{1}{N_k - n_k^c}\Bigg) < \hat{S}(t_{k-1})\left(1 - \frac{1}{N_k}\right) = \hat{S}(t_k | n_k^c = 0)$$

Thus, any censoring in $(t_{k-1}, t_k]$ decreases $\hat{S}(t_k)$. Now, to understand the impact of censoring on subsequent events, we consider a scenario where n_k^c patients are censored in $(t_{k-1}, t_k]$ and there were no more censoring in the interval $(t_k, t_{k+l}]$ until l additional events are observed. In this case,

$$\hat{S}(t_{k+l} | n_k^c > 0) = \hat{S}(t_{k-1}) \quad 1 - \frac{1}{N_k - n_k^c}\Bigg)\left(1 - \frac{1}{N_k - n_k^c - 1}\right)\cdots \quad 1 - \frac{1}{N_k - n_k^c - l}\Bigg)$$

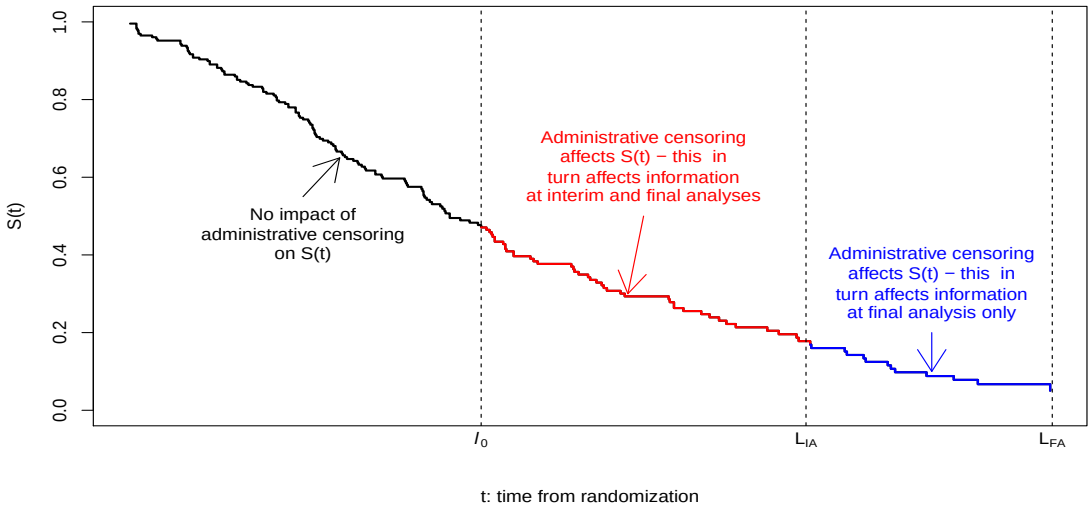

Figure 1: Impact of administrative censoring on K-M curve (l_0 = smallest censoring time, L_{IA} = maximum follow-up time at interim analysis, L_{FA} = maximum follow-up time at final analysis). In absence of any lost to follow-up (LTFU) drop out, $l_0 = L - E$ where L is total study duration, and E is enrollment duration. With LTFU drop out, l_0 can be much smaller than $L - E$.

It is easy to see that

$$\hat{S}(t_{k+l}|n_k^c = 0) - \hat{S}(t_{k+l}|n_k^c > 0) > \hat{S}(t_k|n_k^c = 0) - \hat{S}(t_k|n_k^c > 0) \tag{8}$$

That is, $n_k^c > 0$ causes a greater drop in $\hat{S}(t_{k+l})$ compared to that in $\hat{S}(t_k)$. Thus, each censoring affects $\hat{S}(\cdot)$ at all subsequent events with increasing drops in $\hat{S}(\cdot)$ at later events. Furthermore, any censoring impacts only subsequent events, but not the previous events. These two impacts together imply that the drop in $\hat{S}(\cdot)$ due to censoring is greater at later events compared to earlier events.

Since censoring affects $\hat{S}(\cdot)$, consequently, censoring also affect the information contributed by each event, $\frac{a}{(a+1)^2}[\hat{S}(\cdot)]^{2\rho}[1 - \hat{S}(\cdot)]^{2\gamma}$ (see Eq. (2)). For $\rho = 0$ (i.e., when accounting for late separation), censoring increases the information $\frac{a}{(a+1)^2}[1 - \hat{S}(\cdot)]^{2\gamma}$ accrued by each event. Similarly, censoring decreases the information accrued by each event when accounting for early separation (i.e., $\gamma = 0$). In both the cases, changes in information due to censoring is greater at later events compared to earlier events. The effect of censoring is also controlled by the degree of weighting as determined by ρ and γ with greater weight resulting in more effect of censoring.

3.2.2 Comparison between IF with and without censoring

As before, let l_0 be the smallest censoring time, L_{IA} be the maximum follow-up time at interim analysis and L_{FA} be the maximum follow-up time at final analysis (see Figure 1). For now assume, $l_0 \leq L_{IA}$. Further assume D_0, D_{IA} and D_{FA} events were observed by time l_0, L_{IA} and L_{FA}, respectively. Denote the three disjoint contiguous time intervals as: $T_0 = (0, l_0]$, $T_1 = (l_0, L_{IA}]$ and

$T_2 = (L_{IA}, L_{FA}]$. Further, denote the information accrued in T_0, T_1 and T_2 without censoring as x_0, x_1 and x_2, respectively, and the change in information due to censoring as Δx_1 and Δx_2 in T_1 and T_2, respectively. In other words, x_0 is the information from subjects with event times not exceeding the smallest censoring time, x_1 is the information from subjects with event times greater than the smallest censoring time but not exceeding the event time of $D_{IA}{}^{th}$ event, and x_2 is the information from subjects with event times beyond that of $D_{IA}{}^{th}$ event. Note that censoring does not affect information accrued in T_0. Denoting IF without considering the effect of censoring as $IF_0 \equiv IF_0(\rho, \gamma)$ and with censoring as IF_c, we have

$$\text{IF without censoring: } IF_0 = \frac{x_0 + x_1}{x_0 + x_1 + x_2}$$

$$\text{IF with censoring: } IF_c = \frac{x_0 + x_1 + \Delta x_1}{x_0 + x_1 + x_2 + \Delta x_1 + \Delta x_2}$$

Note that IF does not change due to censoring (that is, $IF_c = IF_0$)) as long as the relative increase in information in $(0, L_{IA}]$ and $(0, L]$ are equal; that is, when

$$\frac{\Delta x_1}{x_0 + x_1} = \frac{\Delta x_1 + \Delta x_2}{x_0 + x_1 + x_2}$$

We can express the difference in IF as

$$IF_0 - IF_c = \left(\frac{\Delta x_1}{x_0 + x_1 + x_2 + \Delta x_1 + \Delta x_2}\right) \cdot [IF_0 \cdot \Delta x_2 - (1 - IF_0) \cdot \Delta x_1] \tag{9}$$

For ease of discussion, consider the weighting for late separation only. Under late separation, censoring reduces the information (see Section 3.2.1) and thus both Δx_1 and Δx_2 are positive. Therefore, both $IF_0 \cdot \Delta x_2$ and $(1 - IF_0)\Delta x_1$ are positive. Hence, IF_c cannot differ much from IF_0. Moreover, the quantity $IF_0 \cdot \Delta x_2$ is larger than $(1 - IF_0)\Delta x_1$ if the following two conditions are satisfied:

(a) $IF_0 > (1 - IF_0)$; or equivalently, $IF_0 > 0.5$, and

(b) $\Delta x_1 < \Delta x_2$, which holds if the number of events in T_2 (i.e., $D_{FA} - D_{IA}$) is greater than the number of events in T_1 (i.e., $D_{IA} - D_0$) (see Eq. (8)).

If the above two conditions are not met, we can not make a definitive statement whether or not $IF_0 \cdot \Delta x_2$ is bigger than $(1 - IF_0) \cdot \Delta x_1$. We note that the above two conditions cannot both fail, as explained in the two points below:

1. Condition (a) may not be satisfied when either $D_{IA}/D_{FA} < 0.5$ is small implying that only relatively fewer events contribute to x_1, or γ is relatively large (i.e., steep increase in weight towards later events), implying that $\Delta x_1 << \Delta x_2$ favouring $IF_0 \cdot \Delta x_2$ to be greater than $(1 - IF_0)\Delta x_1$.

2. Condition (b) may not be satisfied if D_{IA}/D_{FA} is close to 1 or γ is relatively small. A value of D_{IA}/D_{FA} closer to 1 implies $IF_0 >> (1 - IF_0)$, again favoring $IF_0 \cdot \Delta x_2$ to be greater than $(1 - IF_0)\Delta x_1$, whereas a smaller value of γ limits the overall impact of censoring (see Section 3.2.1).

Therefore, the quantity $IF_0 \cdot \Delta x_2$ would either exceed or be very close to $(1 - IF_0)\Delta x_1$ in most practical cases. Consequently, under late separation, IF_c is likely to be smaller than IF_0 under almost every reasonable censoring. In fact, it would be challenging to construct a censoring pattern that would lead IF_c greater than IF_0. Therefore, we conclude that $IF_0 \gtrapprox IF_c$ under late separation. Recall at the beginning of this sub-section, we assumed $l_0 \leq L_{IA}$. Now if $l_0 > L_{IA}$, then clearly, $\Delta x_1 = 0$ implying $IF_c < IF_0$, since $\Delta x_2 > 0$ under late censoring. Hence, our conclusion that $IF_0 \gtrapprox IF_c$ under late separation is well justified.

Similarly, under early separation, both Δx_1 and Δx_2 are negative quantities, and hence, $IF_0 \lessapprox IF_c$. In summary, these are the major takeaway lessons from the above discussion:

1. Usually, censoring works in the same direction of the weight function when accounting for either late or early separation. That is, censoring further reduces (increases) IF when accounting for late (early) separation. Therefore, IF_0 represents the maximum (minimum) IF under late (early) separation.

2. IF with censoring (IF_c) cannot differ much from IF without censoring (IF_0).

3. The following relationship holds when accounting for late separation

IF for the SLR test (D_{IA}/D_{FA})	$>$	IF for FH test without censoring (IF_0)	$\gtrapprox$	IF for FH test with censoring (IF_c)

4. The following relationship holds when accounting for early separation

IF for SLR test (D_{IA}/D_{FA})	$<$	IF for FH test without censoring (IF_0)	$\lessapprox$	IF for FH test with censoring (IF_c)

Thus ignoring the effect of censoring might inflate the probability of type I error under late separation and may compromise the interim power under early separation.

3.2.3 Estimation of information fraction with censoring

The realization of a censoring pattern in a study is a convolution of three distinct distributions—enrollment, LTFU and TTE distributions. Since censoring affects IF, a precise knowledge of these three distributions is necessary to calculate IF [18–21]. However, since these distributions cannot be predicted precisely at the design stage, an obvious strategy would be to evaluate all possible censoring scenarios and then to choose the minimum IF to preserve the overall size. Note that, censoring ultimately affects $\hat{S}(\cdot)$ (and therefore, the FH weight and IF) through the distribution of number of censored patients (n_k^c) within each time interval $(t_{k-1}, t_k]$ only (see Eq. (4)). We call this "censoring distribution against events" or CDE. Precisely speaking, CDE is defined as distribution of n_k^c at each observed event k and n_k^c is defined as follows

$$n_k^c = \sum_{i=1}^{N} I(i^{th} \text{ subject is censored}) \times I(t_{k-1} < \text{follow-up time of } i^{th} \text{ subject} \leq t_k)$$

where $I(\cdot)$ is the indicator function. Since it is relatively simpler to work with a single distribution than to work with three different distributions, we estimate the IF_c using CDE. For a given CDE,

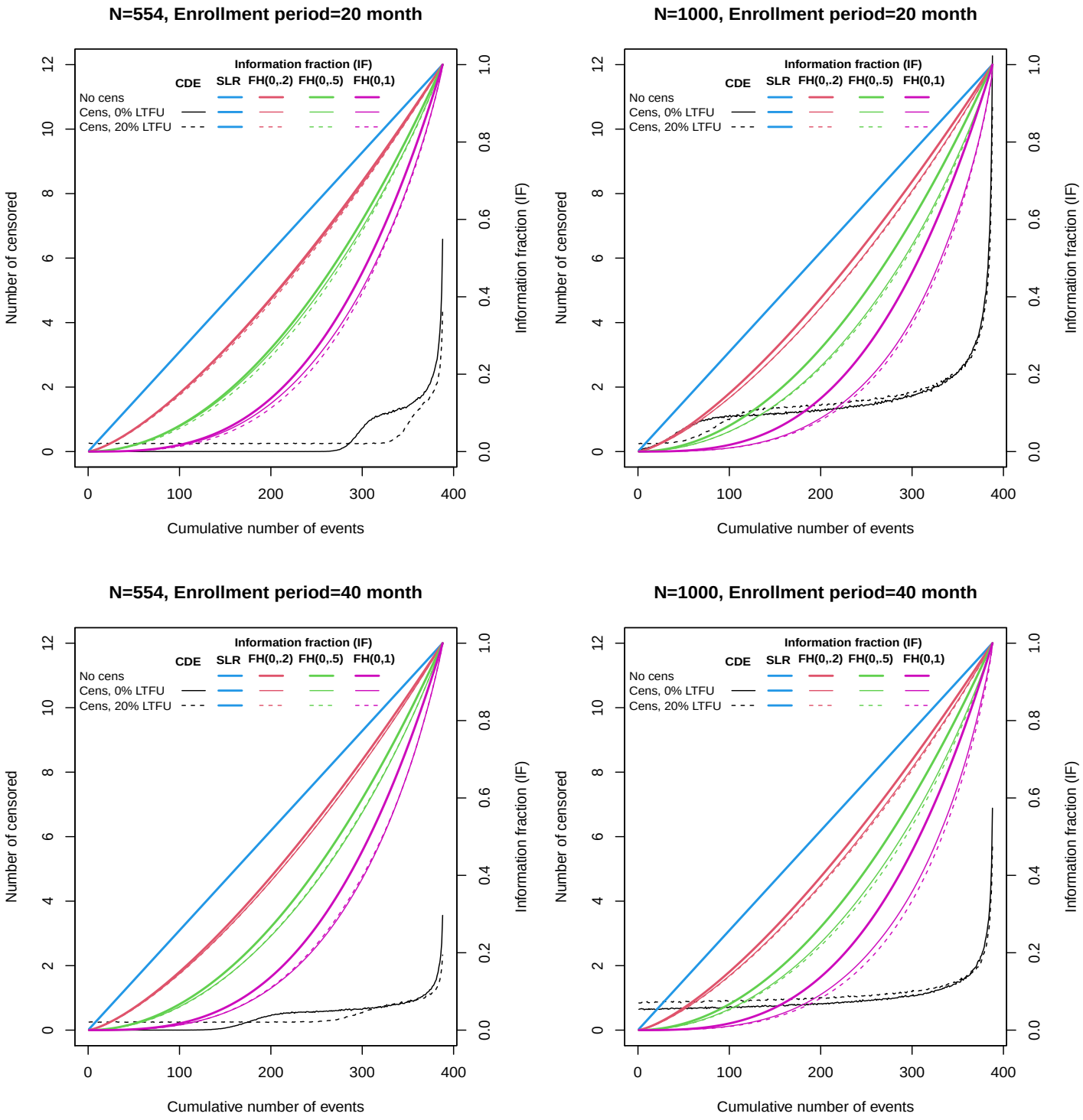

Figure 2: Average CDE and minimum IF (i.e., IF_c) for FH test and SLR test with 388 events based on 10,000 simulated data sets as described in Section 3.2.3. The black lines (representing CDE) should be compared with the first y-axis and all other lines (representing IF) should be compared with the second y-axis. IF for SLR test is the ratio of events and is not impacted by censoring. IF for FH tests without censoring are obtained using Eq. (7). Clearly, IF for FH test is smaller than IF for SLR test. Further, censoring reduces IF for FH tests with increased effect of censoring as weight in FH test increases.

IF_c (i.e., IF with censoring) can be computed by plugging n_k^c in $\hat{S}(t_k)$ in Eq. (4) and then $\hat{S}(t_k)$ in Eq. (3). `R` codes to calculate CDE and IF_c are presented in Appendix 2 and 4, respectively.

Examples of CDEs and resultant IFs are displayed in Figure 2 with uniform enrollment distribution over an enrollment period of 20 months or 40 months, exponential TTE distribution (with a median of 14.1 months) and exponential LTFU distribution (with an overall LTFU rate of either 0% or 20%). The total sample size was varied between 554 and 1000 with 388 final events. For each scenario, 10,000 simulated data sets were generated, CDE and IF_c for FH tests were obtained for each data set and the average CDE and the minimum IF are presented. As can be seen in Figure 2, a larger sample size results in a relatively more administratively censored patients leading to more left-skewed CDE. Prolongation of enrollment period can have a similar impact as larger sample size; however, excessive prolongation of enrollment period will have a reverse effect as the targeted number of events for final analysis may be reached before the completion of enrollment period thereby reducing the total number of administratively censored patients. A higher proportion of LTFU patients tends to distribute the censored patients evenly over the events compared to the no LTFU scenario, in which case CDE tends to be relatively more left-skewed. The distribution of LTFU with a fixed overall LTFU rate has very limited impact on CDE when compared to a uniform LTFU distribution (not shown in the figure).

Figure 2 also displays the IF for SLR test and three FH tests (FH(0,0.2), FH(0, 0.5), and FH(0, 1.0)) accounting for late separation. Clearly, IFs for all three FH tests are much smaller than IF for SLR test (i.e., ratio of events). Further, for each of the FH tests, IF with censoring (IF_c) is smaller than that without censoring (IF_0). As the γ in FH(ρ, γ) increases, (a) both the IF(SLR) $-IF_0$ and $IF_0 - IF_c$ increase, and (b) the effect of censoring ($IF_0 - IF_c$) is substantially smaller than the effect of weighting alone (i.e., IF(SLR) $-IF_0$). For the FH test accounting for early separation, the effect would be almost similar, though in the reverse direction; that is, an increase in IF due to weighting and censoring with increased effect of censoring as weighting increases.

A particular CDE is a realization of specified enrollment, TTE and LTFU distributions. Therefore, we must evaluate an exhaustive list of all possible CDEs to determine the minimum IF_c. For that, we first take note of the following three distinct features in CDE plots presented in Figure 2:

F1: Initially, the CDE plot remains flat and very close to zero. At this stage, administrative censoring did not begin and censoring comes primarily from LTFU.

F2: Thereafter, censoring increases with time. At this stage, aside from LTFU, effects of administratively censored patients become more visible.

F3: There is a sharp increase in censoring towards the end because of a relatively large number of administratively censored subjects.

We propose to mimic the above features in CDE and then to characterize all potential CDEs by varying the timing of onset of these stages and the distribution within each stage consistent with the above observations. We have illustrated this in Section 4 (see Eq. (10)).

3.2.4 Factors influencing the effect of censoring on information fraction

The magnitude of contribution to IF due to censoring depends on several factors. We discuss the impact of these factors only in the context of the FH test accounting for late separation, leaving other contexts to the reader.

1) Number of patients censored: Quite naturally, the impact of censoring on IF will be higher when more patients are censored. Enrollment of relatively large number of patients to reach a target number of events in a short duration implies that large number patients would be censored resulting in a greater impact of censoring on IF. In this context, we also note that, under late separation, attempting to shorten the trial duration by increasing the total sample size is not desirable, since the data may not be mature enough to manifest late separation and the overall power of the test may be compromised.

2) Proportion of events at interim analysis: If the number of interim events are smaller than the events at the final analysis, then impact of censoring on information accrued by interim events could be very small compared to the that on post-interim events causing a greater effect of censoring on IF. On the contrary, we can limit the effect of censoring on IF by conducting the interim analysis closer to the final analysis.

3) Weight function: The quantity $IF_0 - IF_c$ is largely impacted by the quantity $IF_0 \cdot \Delta x_2$ relative to $(1 - IF_0) \cdot \Delta x_1$ (see Eq. (9)). As the imbalance in weight between earlier and later events increases, Δx_2 also increases, resulting in smaller IF_c.

4) Shape of the CDEs: A large number of patients censored immediately before an interim analysis hardly impacts information accrued at interim analysis whereas the effects of such censoring is fully realized through post-interim events. Such a censoring pattern causes a greater decrease in IF when accounting for late separation.

4 Example

The Set-up

To illustrate the proposed method, we consider a clinical trial design in which we intend to observe 388 events at the time of final analysis. We assume an overall treatment-to-control HR of 0.75. Due to a potential delayed treatment effect, we wish to consider the FH test for the test of HR with a one-sided significance level of 0.025. The planned total sample size was 554; and these patients will be assigned to the two treatment arms in the ratio 1:1. We consider only one planned interim analysis with timing of interim analysis varied from 291 events (75% of target) to 194 events (50% of target). The two treatment arms will be compared using the FH test accounting for late separation: Hence, we set $\rho = 0$ and $\gamma > 0$ (where a larger γ represents more imbalance in weights in favor of later events). The overall probability of type I error of testing will be split between interim and final analysis according to Lan-DeMets O'brien-Fleming approximate spending function. We further assume that patients will be enrolled over a period of 20 months (i.e., $E = 20$), and the total study duration is expected to be 33 months (i.e., $L = 33$) based on an assumed control median of 11 months.

Determining IF_0 (i.e., IF accounting for only variable weighting, but no-censoring)

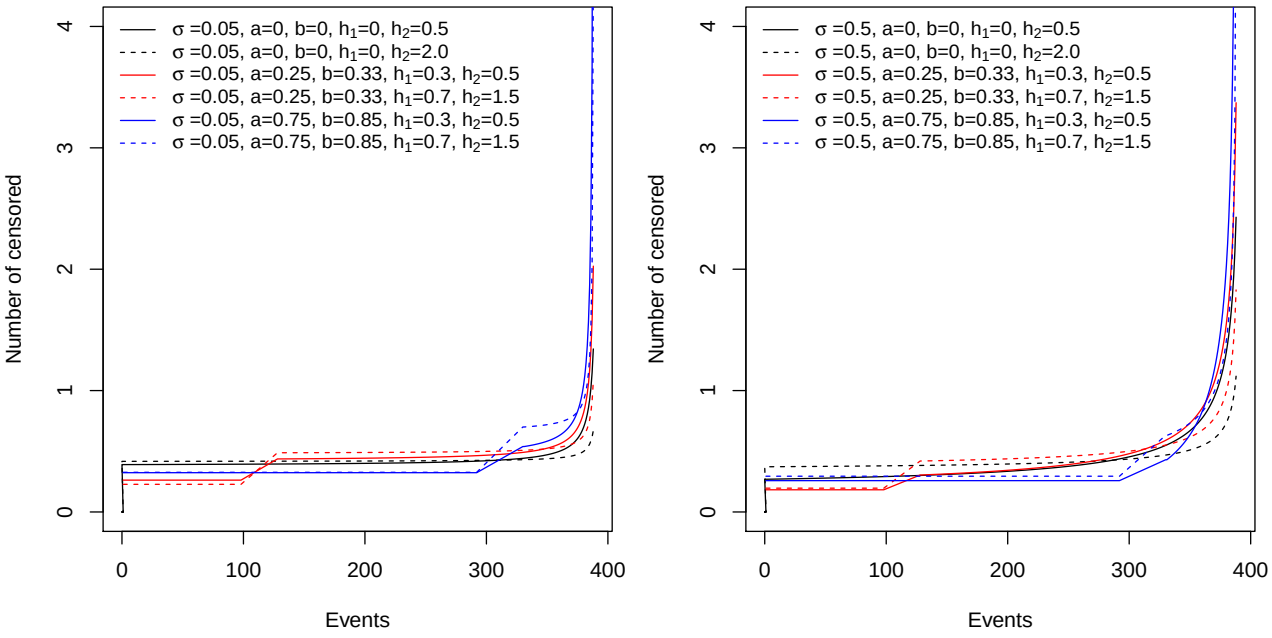

Figure 3: Some CDEs generated based on Eq. (10) and later used in calculation of IF_c (i.e., IF with censoring). The y-axis is truncated at 4 in an attempt to make the features in CDE relatively more visible. In fact, an exhaustive list of CDEs were generated by varying σ, a, b, and h_1 and h_2; however, due to space constraints only a handful of CDEs are displayed as examples.

Under late separation, IF_0 can be directly obtained using the number of events at interim and final analyses using Eq. (7). Table 1 summarizes IF_0 obtained for different weight functions $FH(\rho = 0, \gamma > 0)$, as γ ranges over [0, 1]. $FH(\rho = 0, \gamma = 0))$ represents SLR test and hence, IF_0 equals to the ratio of interim to final counts of events (i.e., $IF_0 = 0.75$ at 75% events and $IF_0 = 0.50$ at 50% events). For $\gamma > 0$, we see a drop in IF_0 compared to that of SLR. This drop becomes larger as the γ (i.e., difference in weights in favor of later events) increases due to slow accrual of information.

Estimating IF_c (i.e., IF accounting for both variable weighting and censoring)

As explained in Section 3.2.3, we evaluated the effect of censoring through CDEs (i.e., distribution of censored patients between any two successive events). In order to approximate CDE, we have first considered an underlying functional form $f(x), x \in (0, 1]$ representing the features (F1–F3) of CDEs observed in Section 3.2.3 as follows:

$$f(x) \propto \begin{cases} h_1 & 0 < x < a \quad \text{(corresponds to feature F1)} \\ h_1 + (h_2 - h_1) \cdot x/b & a \leq x < b \quad \text{(corresponds to feature F2)} \\ h_2 + g\,(1 - x|\sigma) - g\,(1 - b|\sigma) & b \leq x \leq 1 \quad \text{(corresponds to feature F3)} \end{cases} \tag{10}$$

where $0 \leq h_1 \leq h_2$ and $g(x|\sigma)$ is the density of a log-logistic distribution with shape σ $(0 < \sigma \leq 1)$

Table 1: IF without censoring (IF_0), minimum IF with censoring (IF_c) and associated α rejection boundaries for the FH($\rho = 0, \gamma$) test at interim and final analysis with an overall one-sided probability of type I error of 2.5%. The column $\gamma = 0$ represents the standard log-rank test.

Fleming Harrington test	Accounting for variable weighting only (i.e., without censoring)			Accounting for both variable weighting and censoring		
	Information fraction (%)	α (1-sided)		Information fraction (%)	α (1-sided)	
		Interim	Final		Interim	Final
Interim analysis at 75% event						
FH($\rho = 0, \gamma = 0$)	75.00	0.0096	0.0221	75.00	0.0096	0.0221
FH($\rho = 0, \gamma = 0.1$)	72.17	0.0083	0.0225	70.14	0.0074	0.0227
FH($\rho = 0, \gamma = 0.25$)	66.53	0.0060	0.0232	63.14	0.0048	0.0235
FH($\rho = 0, \gamma = 0.5$)	58.01	0.0033	0.0240	52.46	0.0020	0.0244
FH($\rho = 0, \gamma = 0.75$)	50.70	0.0016	0.0245	43.13	0.0006	0.0248
FH($\rho = 0, \gamma = 1$)	44.26	0.0008	0.0247	35.09	0.0002	0.0249
Interim analysis at 50% event						
FH($\rho = 0, \gamma = 0$)	50.00	0.0015	0.0245	50.00	0.0015	0.0245
FH($\rho = 0, \gamma = 0.1$)	44.40	0.0008	0.0247	42.90	0.0006	0.0248
FH($\rho = 0, \gamma = 0.25$)	36.26	0.0002	0.0249	33.84	0.0001	0.0250
FH($\rho = 0, \gamma = 0.5$)	25.86	0.0000	0.0250	22.47	0.0000	0.0250
FH($\rho = 0, \gamma = 0.75$)	18.44	0.0000	0.0250	14.73	0.0000	0.0250
FH($\rho = 0, \gamma = 1$)	13.15	0.0000	0.0250	9.54	0.0000	0.0250

Interim and final α are determined using the gsDesign package as follows:
`gsDesign(k=2, sfu=sfLDOF, test.type=1, alpha=0.025, timing=c(IF, 1))|`

and scale 1. Subsequently, n_k^c for CDE were obtained as follows:

$$n_k^c = f\left(\frac{k}{D_{FA}}\right) \cdot (N - D_{FA}) / \sum_{r=1}^{D_{FA}} f\left(\frac{r}{D_{FA}}\right) \tag{11}$$

Note that n_k^c may be fraction representing expected number of patients censored in consistent with CDEs observed in Figure 2. The `R` codes to generate CDE based on Eq. (10) and (11) is presented in Appendix 3. In this example, we have varied σ, a, b, and h_1, h_2 as follows to generate a wide range of CDEs to evaluate: $\sigma = 0.001(.15)0.901$ and 1; a=0(.15).90 and 1; $b = a(.15)0.90$ and 1; $h_1 = 0(0.5)5$; $h_2 = h_1(0.5)5$. Some of the CDEs are plotted in Figure 3. The IF_c were calculated for all CDEs described in Section 3.2.3, and only the minimum IF_c over all CDEs are presented in Table 1. As argued in Section 3.2.2, IF_c is smaller than IF_0 across all values of γ. Generally, IF_c is close to IF_0 for small γ; however, the gap increases with the increase in γ.

Determining rejection boundaries based on minimum IF_c

Interim and final alpha rejection boundaries using IF_0 and the least IF_c are also presented in Table 1. When comparing IF_c with IF_0, it is clear that IF of the FH tests is significantly impacted by the weight function; but the additional impact of censoring is relatively smaller. For example, at the interim analysis with 75% events, the weight $FH(\rho = 0, \gamma = 0.25)$ alone reduces the IF by 8.47% (from 75% to 66.53%). Thereafter, censoring reduces the IF further by 3.39% (from 66.53% to 63.14%), which translates into a difference of only 0.0012 in the interim alpha boundary.

5 Simulation

We conducted a simulation study to evaluate the operating characteristics (size, power under PH assumption and power under specified delayed effects) of the study design presented in Section 4. In all cases, event times on the control arm were generated from an exponential distribution with a median of 11 months. An enrollment period of 20 months was allowed; and within each month, the enrollment times were generated from a uniform distribution. Each patient was assigned to either the treatment or the control arm with probability 1/2 each. To account for LTFU, drop-out times were generated randomly from an exponential distribution with 10% drop-out rate by 3 years.

Denoting event time by T^* and censoring time by C, the follow-up time was determined as $T = \min(T^*, C)$, and the event indicator as $\delta = I(T^* \leq C)$: $\delta = 1$ when event is observed by the time of analysis; otherwise, $\delta = 0$ when patient is either dropped-out or administratively censored. Empirical size and power are reported based on proportion of rejections at interim and final analyses using the α rejection boundaries shown in Table 1.

5.1 Size of the FH test

To evaluate the overall size of the test, survival data on the treatment arm were also generated from an exponential distribution with median time the same as on control arm (that is, 11 months). Results based on 100,000 replicates are presented in Table 2. In Table 2, under both the no censoring scenario and in presence of censoring, we see that the size of the FH test, with α determined based on the corrected IF, is very close to nominal size of 2.5%, and comparable to the size of the SLR (i.e.,

Table 2: Empirical size of Fleming Harrington test with 0.025 nominal type I error (1-sided)

	Simulated with no censoring				Simulated with censoring			
	FH test with α using correct IF		FH test with α using 75% IF		FH test with α using correct IF		FH test with α using 75% IF	
Fleming Harrington test	Interim size	Overall size	Interim size	Overalll size	Interim size	Overal size	Interim size	Overall size
FH($\rho = 0, \gamma = 0$)	0.0093	0.0241	0.0093	0.0241	0.0100	0.0255	0.0100	0.0255
FH($\rho = 0, \gamma = 0.1$)	0.0079	0.0244	0.0093	0.0248	0.0079	0.0255	0.0100	0.0260
FH($\rho = 0, \gamma = 0.25$)	0.0057	0.0248	0.0096	0.0258	0.0050	0.0251	0.0102	0.0265
FH($\rho = 0, \gamma = 0.5$)	0.0034	0.0252	0.0096	0.0268	0.0021	0.0252	0.0100	0.0271
FH($\rho = 0, \gamma = 0.75$)	0.0016	0.0251	0.0098	0.0274	0.0007	0.0253	0.0098	0.0277
FH($\rho = 0, \gamma = 1$)	0.0008	0.0252	0.0101	0.0278	0.0002	0.0256	0.0101	0.0281

Table 3: Empirical power of Fleming-Harrington test

	Under PH model		Under late separation	
Fleming Harrington test	Interim power (%)	Overall power (%)	Interim power (%)	Overall power (%)
FH($\rho = 0, \gamma = 0$)	54.36	79.94	33.89	76.69
FH($\rho = 0, \gamma = 0.1$)	50.05	79.71	31.52	79.00
FH($\rho = 0, \gamma = 0.25$)	42.57	78.74	27.93	81.83
FH($\rho = 0, \gamma = 0.5$)	28.18	75.53	20.55	84.28
FH($\rho = 0, \gamma = 0.75$)	14.88	71.96	12.49	85.26
FH($\rho = 0, \gamma = 1$)	7.28	67.98	7.42	85.47

FH(0,0)) test. On the contrary, if the α boundaries based on 75% IF are used for the FH($\rho = 0, \gamma > 0$) testing, we observe a relatively higher rejection at interim analysis causing the overall size to exceed the nominal size, and this inflation in size increases as γ increases. If we continue to use the α rejection boundaries with 75% IF in censoring case as well, the inflation in size worsens as censoring further reduces IF. However, even under censoring, the FH test based on a correct IF can constrain the overall size of the FH test. This finding validates the proposition to use an appropriate IF for $FH(\rho = 0, \gamma > 0)$ test to control overall size of test, and to carry out the FH testing accordingly.

5.2 Power under proportional hazards scenario

Under PH model, the event times in the treatment arm were generated from an exponential distribution with a median of 14.67 months, corresponding to a HR of 75%, as assumed in Section 4. Results based on 20,000 replicates are presented in Table 3, with α boundaries (under "With censoring") presented in Table 1. As expected, the power under the SLR test (i.e., FH($\rho = 0, \gamma = 0$)) is close to the nominal power of 80%, and the power reduces gradually as the imbalance in weights in favor of later events increases gradually (i.e., as γ increases).

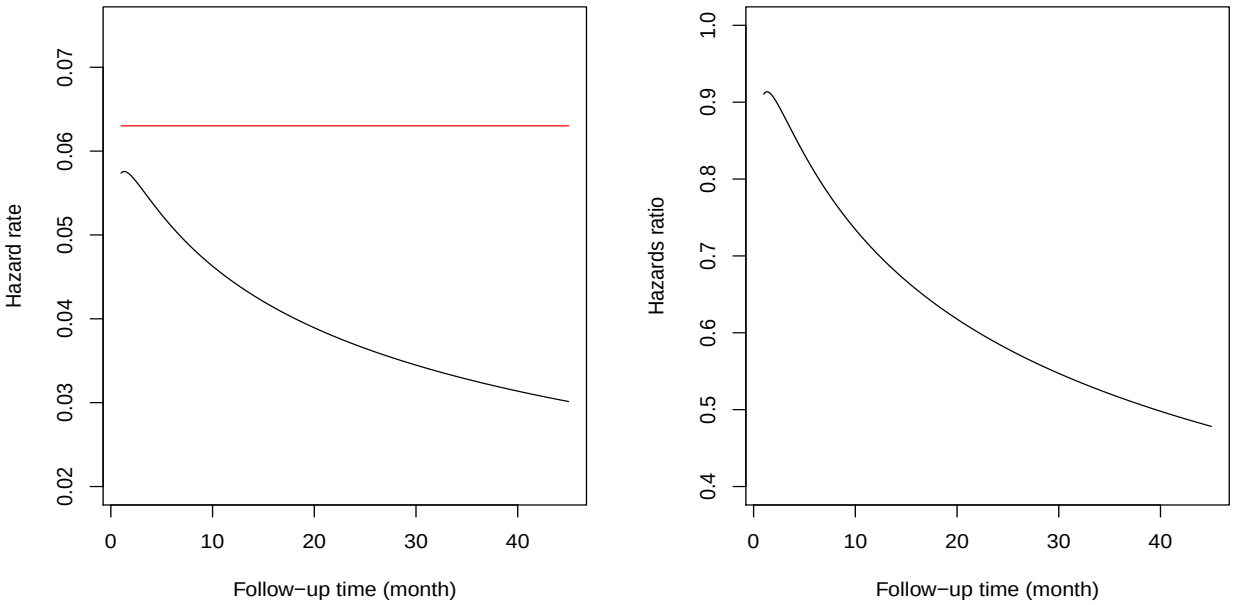

Figure 4: Hazard rate functions of the treatment (black) and the control (red) arms, and the HR (treatment to control) used for simulation event times under late separation in Section 5.3.

5.3 Power under late separation

Under late separation model, the event times in the treatment arm were generated from the following generalized gamma distribution [26, 27].

$$f_T(t|\mu,\sigma,Q) = \frac{|Q|(Q^{-2})^{Q^{-2}}}{\sigma \cdot t \cdot \Gamma(Q^{-2})} exp\left[Q^{-2} \cdot \{Qw - exp(Qw)\}\right] \quad \text{where,} \quad t = \exp(\mu + \sigma w)$$

Event times from this generalized gamma distribution were generated using the `R` function rgengamma() in the flexsurv package [27] with parameters $\mu = 2.9$, $\sigma = 1.271$ and $Q = 0.61$ representing decreasing hazard rate (see Figure 4). With constant hazard rate in the control arm, this also corresponds to late separation characterized by a reduction in HR over time. These parameters were chosen to ensure an estimated HR of 0.70 under the Cox PH model based on 1 million simulated data points in each arm.

Summary of results based on $20,000$ replicates are presented in Table 3 with α boundaries (under "With censoring") presented in Table 1. As expected, the overall power of the SLR test (i.e., FH($\rho = 0, \gamma = 0$)) is down to 76.69%. The power increases gradually as γ increases; however, the improvement beyond $\gamma = 0.75$ is very minimal. Therefore, some of the power lost due to late separation can be recovered by putting more weights to the later events. As γ increases, the power of the interim test decreases because of the combined effects of (a) reduced IF at interim which translates to tighter rejection boundaries with FH($\rho = 0, \gamma > 0$), and (b) only a partial realization of the late separation.

6 Implementation of FH test in a group-sequential clinical trial design

We now summarize the overall strategy of implementing the FH test in a group-sequential clinical trial with a total sample size of N, the number of interim events D_{IA} and that at final analysis D_{FA}.

Step 1. Determine, IF_0, the IF under variable weighting without censoring: IF_0 can be determined analytically using Eq. (6) with the knowledge of N, D_{IA} and D_{FA}. When accounting for delayed effects, calculation of IF_0 needs only D_{IA} and D_{FA} (see Eq. (7)).

Step 2. Determine the minimum IF_c, the IF with censoring: IF_c should be determined under various plausible CDEs to identify plausible range for IF, as presented in Section 3.2.3. One may adopt the strategy presented in Section 4 to identify various CDEs by including some other plausible scenarios or by dropping some unlikely scenarios. As the interim analysis approaches, we may get to see the actual accrual and drop out patterns, IF_c may be updated by narrowing down the plausible CDEs, and the α-rejection boundaries should be updated accordingly.

Step 3. Determine α rejection boundaries for interim and final analyses using minimum IF_c.

Step 4. Select the weight function $FH(\rho, \gamma)$ via simulation: Exhaustive simulation should be carried out to compare the operating characteristics of FH test with the SLR test under both PH scenario and the expected non-PH scenario. A suitable weight function should be chosen to minimize the power loss under PH scenario and to maximize power gain under expected non-PH scenario.

In Step 3, rejection boundaries are determined solely based on the IF_c. Even though IF_0 is not used further, we strongly recommend determining IF_0 because of its proximity to IF_c.

The above strategy can be easily extended to implement FH test in a group sequential design with multiple interim analyses, say with events $D_{IA_1}, D_{IA_2}, \cdots, D_{IA_M}$. In such a case, in the design stage, one needs to calculate IF_c for each interim analysis using all plausible CDEs. In order to minimize the loss of power, one may update the IF at each interim analysis using all available information (on pooled survival distribution, enrollment distribution and LTFU distribution) and the α-rejection boundaries can be updated accordingly. While updating the α-rejection boundaries based on revised IF_c, one must account for the α already spent in the previous analyses in order to control the overall type-I error of the study.

7 Discussion

Previous researches ([7, 8]) have shown that in a single look design, the FH test preserves the size of the test; it has a minimal loss of power under the PH scenario; and it has a substantial gain in power under the non-PH scenario. These desired properties of the FH test continue to hold under a group-sequential design as long as IF at the interim look is determined correctly, which can be very far from the proportion of interim events. Generally speaking, the IF under a FH test is affected by the two factors: (a) variable weighting of events (as ascertained by the weight function); and (b) censoring. In this paper, we have separated the effects of these two factors and have shown that

these two factors affect IF in the same direction. That is, when accounting for late (early) separation, assigning higher weights towards the later (earlier) events reduces (increases) IF, and censoring on top of that further reduces (increases) this IF. We further concluded that for any given weight function, its effect on IF is much larger than the additional effect of censoring (see Section 3.2.2) which enables us to construct a reasonable range of IF. Subsequently, the minimum IF over a set of plausible non-PH scenarios should be used to preserve the overall type I error.

To account for the impact of weight function on IF, we have derived an analytical formula using total sample size and numbers of events at interim and final analyses. (Moreover, under the special case of late separation, IF does not even depend on the sample size, see Eq. (7)). But to study the censoring effect on IF we have proceeded in a somewhat ad hoc manner. Nonetheless, we have demonstrated that censoring has a relatively much smaller impact on IF. Because the nature of censoring is uncertain at the design stage, we should usually consider all plausible censoring scenarios and re-evaluate the IF as we get closer to the interim analysis, by which time we may acquire a relatively better idea about the enrollment distribution. The knowledge of the enrollment distribution, the projected pooled TTE distributions and, LTFU rate based on the blinded information on the timings of the pooled events may help us rule out some implausible censoring scenarios and reduced the number of plausible scenarios of censoring. This in turn may help to narrow down the range of IF and ultimately avoid being overly conservative at the interim analysis, and increase the interim power.

As mentioned in Section 6, the choice of weight function $FH(\rho, \gamma)$ should be driven by comparative assessments of operating characteristics of the FH test against the SLR test both under the PH scenario and a few anticipated non-PH scenarios. A suitable weight function should be chosen to minimize the power loss under PH scenario, and to maximize the power gain under late separation. As seen in our simulation study, beyond a certain point any additional increase in weight does not enhance the power under late separation; but it can drastically decrease the power under PH scenario. Also, assigning a very high weight to later events to account for delayed effects may severely impact IF and reduce the chance of detecting difference at interim analysis. Hence, in general, we suggest allowing only a modest imbalance in weighting, unless there is a strong justification to do otherwise.

Finally, although in this paper we have mainly focused on the WLR test with FH weighting because of its popularity and flexibility to account for most common types of non-PH scenarios, the principles discussed in this paper can be easily extended to the other WLR tests as well.

Acknowledgement

The authors wish to thank two anonymous referees for several corrections and suggestions that greatly improved the paper.

References

[1] Stupp R, Mason WP, Van Den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352(10):987-96.

[2] Kantarjian HM, DeAngelo DJ, Stelljes M, et al. Inotuzumab ozogamicin versus standard therapy for acute lymphoblastic leukemia. N Engl J Med. 2016;375(8):740-53.

[3] Ferris RL, Blumenschein Jr G, Fayette J, et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N Engl J Med. 2016;375:1856-67.

[4] Public workshop: Oncology clinical trials in the presence of non-proportional hazards. US Food and Drug Administration Web site. https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards. Published 2018. Accessed Dec 26, 2020.

[5] Hoos A. Evolution of end points for cancer immunotherapy trials. Ann Oncol. 2012;23(suppl_8):viii47-52.

[6] Saad ED, Zalcberg JR, Péron J, Coart E, Burzykowski T, Buyse M. Understanding and communicating measures of treatment effect on survival: can we do better?. JNCI: J Natl Cancer Inst. 2018;110(3):232-40.

[7] Fine GD. Consequences of delayed treatment effects on analysis of time-to-event endpoints. Drug Inf J. 2007;41(4):535-9.

[8] Su Z, Zhu M. Is it time for the weighted log-rank test to play a more important role in confirmatory trials? Contemporary Clinical Trials Communications. 2018 Jun;10:A1-A2.

[9] Zhang J, Pulkstenis E. Sample size and power of survival trials in group sequential design with delayed treatment effect. Stat Biopharm Res. 2016;8(3):268-75.

[10] Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics. 2010;66(1):30-8.

[11] Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. Biometrika. 1990;77(4):853-64.

[12] Xu Z, Zhen B, Park Y, Zhu B. Designing therapeutic cancer vaccine trials with delayed treatment effect. Stat Med. 2017;36(4):592-605.

[13] Fleming TR, Harrington DP. Counting Processes and Survival Analysis. John Wiley, New York; 1991.

[14] Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982;69(3):553-66.

[15] Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. J Am Stat Assoc. 1982;77(380):855-61.

[16] Magirr D, Burman CF. Modestly weighted logrank tests. Stat Med. 2019;38(20):3782-90.

[17] Lan KKG, Zucker DM. Sequential monitoring of clinical trials: the role of information and brownian motion. Stat Med. 1993; 12:753–765.

[18] Gillen DL, Emerson SS. Information growth in a family of weighted logrank statistics under repeated analyses. Seq Anal. 2005;24(1):1-22.

[19] Brummel SS, Gillen DL. Flexibly monitoring group sequential survival trials when testing is based upon a weighted log-rank statistic. Seq Anal. 2014;33(1):39-59.

[20] Lan KG, Rosenberger WF, Lachin JM. Sequential monitoring of survival data with the Wilcoxon statistic. Biometrics. 1995;51(3):1175-83.

[21] Hasegawa T. Group sequential monitoring based on the weighted log-rank test statistic with the Fleming–Harrington class of weights in cancer vaccine studies. Pharm Stat. 2016;15(5):412-19.

[22] Jiménez JL, Stalbovskaya V, Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. Pharm Stat. 2019;18(3):287-303.

[23] Kundu MG. Comments on "Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects" by Jose Jimenez, Viktoriya Stalbovskaya, and Byron Jones. Pharm Stat. 18:287-303, 2019, DOI: 10.1002/pst.1923. Pharm Stat. 2020; 19(5): 733-735.

[24] Lan KG, Wittes J. The B-value: a tool for monitoring data. Biometrics. 1988;44(2):579-85.

[25] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659–663.

[26] Stacy EW. A generalization of the gamma distribution. Ann Math Stat. 1962;33(3):1187-92.

[27] Jackson C. flexsurv: Flexible parametric survival and multi-state models. R package version 1.1.1. https://cran.r-project.org/web/packages/flexsurv/flexsurv.pdf. Published Mar 2019. Accessed Mar 12, 2020.

Appendix

Appendix 1: R code to determine IF without censoring (IF_0) for a given weight of $FH(\rho, \gamma)$

```
IF0<- function(D.FA, D.IA, rho=0, gamma=0, N=NULL){
  if(rho>0){
     if (!is.numeric(N)) stop('Please specify N when rho is greater than 0\n')
     x1<- 1:D.IA; x2<- 1:D.FA
     IF<- sum((N-x1)^(2*rho)*x1^(2*gamma))/sum((N-x2)^(2*rho)*x2^(2*gamma))
  }
  else if(rho==0){
     IF<- sum((1:D.IA)^(2*gamma))/sum((1:D.FA)^(2*gamma))
  }
  IF
}

#--- Examples
IF0(D.FA=388, D.IA=291)  #IF0 under log-rank test
IF0(D.FA=388, D.IA=291, rho=0, gamma=0.25) #IF0 under FH(0, 0.25)
IF0(D.FA=388, D.IA=291, rho=0.25, gamma=0, N=554) #IF0 under FH(0.25, 0)
IF0(D.FA=388, D.IA=291, rho=0.25, gamma=0.25, N=554) #IF0 under FH(0.25, 0.25)
```

Appendix 2: R code to determine CDE under censoring for a given dataset

```
library(survival)
library(rccmisc)
obs.CDE<- function(data, futime, fustat){
T <- data[[futime]]
delta <- data[[fustat]]
survfit.out<- summary(survfit(Surv(T, delta) ~ 1))
D= sum(delta)
tab<- data.frame(n.events= cumsum(survfit.out$n.event),
                 new.censored=survfit.out$n.censor)
tab.dummy<- data.frame(n.events=1:D, tot.censored=rep(0,D))
tab<- merge(tab.dummy, tab, by="n.events", all=TRUE)
tab<- transform(tab, n.censored=psum(tot.censored, new.censored, na.rm=T))
tab[,c("n.events", "n.censored")]
}

#--- Examples
obs.CDE(data=ovarian, futime="futime", fustat="fustat")
```

Appendix 3: R code to simulate CDE based on Equation (10)

```
library(actuar)
sim.CDE<- function(N, D, sigma, a=0, b=0, h1=0, h2=0){
N.censor=N-D
if(h2>=h1){   #--- pre-requisite condition
   a1<- round(a*D); b1<- round(b*D); c1<- D #--- scaling to events
   x1<- 1:a1; y1<- rep(h1, length=length(x1))
   x2<- (a1+1):(b1); y2<- seq(h1, h2, length=length(x2))
   x3<- (b1+1):c1; x3.mod<- 1-seq(b,1, length=length(x3)+1)
   y3<- h2+dllogis(x3.mod[1:length(x3)], shape=sigma, scale=1) -
        dllogis(1-b, shape=sigma, scale=1)
   data.frame(n.events=c(x1, x2, x3),
              n.censored=c(y1, y2, y3)*N.censor/sum(y1, y2, y3))
   }
}

#--- Examples
res<- sim.CDE(N=554, D=388, sigma=0.05, a=0.25, b=0.33, h1=0.1, h2=0.5)
plot(res$n.events, res$n.censored, type='l', col=1, lty=1)
```

Appendix 4: R code to determine IF with censoring (IF_c) for a given CDE and a given weight of $FH(\rho,\gamma)$

```
IFc<- function(cde, N, D.IA, D.FA, rho=0, gamma=0){
St<- function(prev.St, n.risk, new.event) prev.St*(1-new.event/n.risk)
tab<- cde
n.risk<- N - cumsum(tab$n.censored) - tab$n.events
tab$n.risk<- c(N, n.risk[-c(length(n.risk))])
tab$surv<- NA
tab$surv[1]<- 1-1/N
for(d in 2:nrow(tab)) tab$surv[d]<- St(tab$surv[d-1], tab$n.risk[d], 1)
tab$wt<- (tab$surv^rho)*((1-tab$surv)^gamma)
tab$info<- tab$wt^2
sum(tab$info[1:D.IA])/sum(tab$info[1:D.FA])
}

#--- Examples
library(survival) #-- for ovarian dataset
IFc(cde=obs.CDE(data=ovarian, futime="futime", fustat="fustat"),
    N=nrow(ovarian), D.IA=9, D.FA=12, rho=0, gamma=0.25)
IFc(cde=sim.CDE(N=554, D=388, sigma=0.05, a=0.25, b=0.33, h1=0.1, h2=0.5),
    N=554, D.IA=291, D.FA=388, rho=0, gamma=0.25)
```