International Journal of
*Environmental Research and Public Health*

MDPI

*Article*

# The Inverse Log-Rank Test: A Versatile Procedure for Late Separating Survival Curves

**Jimmy T. Efird** [1,2]

[1] VA Cooperative Studies Program Coordinating Center, Boston, MA 02111, USA; jimmy.efird@stanfordalumni.org
[2] School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA

**Abstract:** Often in the planning phase of a clinical trial, a researcher will need to choose between a standard versus weighted log-rank test (LRT) for investigating right-censored survival data. While a standard LRT is optimal for analyzing evenly distributed but distinct survival events (proportional hazards), an appropriately weighted LRT test may be better suited for handling non-proportional, delayed treatment effects. The "a priori" misspecification of this alternative may result in a substantial loss of power when determining the effectiveness of an experimental drug. In this paper, the standard unweighted and inverse log-rank tests (iLRTs) are compared with the multiple weight, default Max-Combo procedure for analyzing differential late survival outcomes. Unlike combination LRTs that depend on the arbitrary selection of weights, the iLRT by definition is a single weight test and does not require implicit multiplicity correction. Empirically, both weighted methods have reasonable flexibility for assessing continuous survival curve differences from the onset of a study. However, the iLRT may be preferable for accommodating delayed separating survival curves, especially when one arm finishes first. Using standard large-sample methods, the power and sample size for the iLRT are easily estimated without resorting to complex and timely simulations.

**Keywords:** inverse log-rank test; clinical trials; survival analysis; non-proportional hazards; delayed treatment effects



**Citation:** Efird, J.T. The Inverse Log-Rank Test: A Versatile Procedure for Late Separating Survival Curves. *Int. J. Environ. Res. Public Health* **2023**, *20*, 7164. https://doi.org/10.3390/ijerph20247164

Academic Editor: Paul B. Tchounwou

Received: 3 November 2023
Revised: 23 November 2023
Accepted: 6 December 2023
Published: 11 December 2023



## 1. Introduction

Delayed treatment effects are the most common type of non-proportional hazards arising in clinical trials, most notably for immunologic cancer drugs [1–4]. A certain period of exposure may be necessary before achieving a treatment response, owing to the mechanism of action for compounds like PD-1 or PD-L1 inhibitors. A small insignificant difference between survival curves typically is observed initially, or, in some cases, the curves may even cross-over up to a certain time point. Thereafter, the curves diverge and late separation occurs, manifesting a differential treatment effect. While a standard log-rank test (LRT) will remain valid for rejecting the null hypothesis of no survival difference and will control the Type I error rate, the procedure will not be uniformly most powerful when the hazards for the curves are non-proportional, as is the case in late-separating curves [5]. Importantly, power may not necessarily increase as the sample size becomes larger.

An LRT that assigns greater weight to events occurring later in the trial will be more sensitive to delayed treatment effects [6]. However, in the absence of "a priori" knowledge, finding a combination of weights that is best able to collectively accommodate various survival scenarios has been challenging [7]. Non-proportional hazards owing to differential censoring between treatment groups also poses a concern, especially when the censoring occurs with greater frequency toward the later part of the trial [8,9].

The inverse log-rank test (iLRT) is a computationally simple, single weight procedure that is moderately robust in detecting late occurring survival differences. Yet, this test also performs well under proportional hazards. We provide empirical examples to illustrate

the novelty and versatility of this method in comparison with the multiple weight "Max-Combo" procedure and the combinatoric-based "Split-Range" test.

## 2. Materials and Methods

### 2.1. Preliminaries

#### 2.1.1. Hypergeometric Framework for Survival Time Data

Let time $(t_i) \geq 0$ represent the pooled times in which participants in either Group 1 or Group 2 experience an event, respectively. Consider the layout in Table 1, where

$d_{i1}$ = # of events at $(t_i)$ in Group 1;
$d_{i2}$ = # of events at $(t_i)$ in Group 2;
$d_i = d_{i1} + d_{i2}$;
$R_{i1}$ = # of participants available at $(t_i)$ in Group 1;
$R_{i2}$ = # of participants available at $(t_i)$ in Group 2;
$R_i = R_{i1} + R_{i2}$.

**Table 1.** Events and non-events in the risk set at $(t_i)$ by study group.

| | Group 1 | Group 2 | Total |
|---|---|---|---|
| Event | $d_{i1}$ | $d_{i2} = d_i - d_{i1}$ | $d_i$ |
| Non-event | $R_{i1} - d_{i1}$ | $R_{i2} - d_{i2}$ | $R_i - d_i$ |
| Total | $R_{i1}$ | $R_{i2} = R_i - R_{i1}$ | $R_i$ |

Under the null hypothesis that the sets of times in the two groups are equivalent, it follows that $(d_{i1})$, conditional on the marginal total $(d_i)$, has a hypergeometric distribution [10–12]. Consisting of the sum of $(R_{i1})$ Bernoulli trials, each with a mean of $\left(\frac{d_i}{R_i}\right)$, the hypergeometric distribution is written as [13,14]

$$P(X = x_i |\, R_i, d_i, R_{i1}) = P(X = x_i |\, R_i, R_{i1}, d_i) = \frac{\binom{R_{i1}}{x_i}\binom{R_i - R_{i1}}{d_i - x_i}}{\binom{R_i}{d_i}} = \frac{\binom{d_i}{x_i}\binom{R_i - d_i}{R_{i1} - x_i}}{\binom{R_i}{R_{i1}}}, \quad (1)$$

where the random variable $(x_i)$ denotes the number of events $(d_{i1})$ in Group 1 at each time point $(t_i)$. In many applied examples, the events of interest are deaths $(d)$. The value for this variable is bounded below by $max[0,\ R_{i1} - (R_i - d_i)]$ and above by $min[d_i,\ R_{i1}]$. Given equal survival times, the probability of an event occurring at $(t_i)$ is not contingent upon the group to which a patient belongs [15].

Observing that $(x_i)$ is less than or equal to $(R_{i1})$, the number at risk in Group 1 at $(t_i)$, it follows that [16]

$$P(X = x_i |\, R_i, R_{i1}, d_i) = \frac{\prod_{j=0}^{x_i - 1}(d_i - j)\prod_{j=0}^{R_{i1} - x_i - 1}(R_i - d_i - j)}{\prod_{j=0}^{R_{i1} - 1}(R_i - j)}. \quad (2)$$

#### 2.1.2. Expectation and Variance

The properties of the hypergeometric distribution are well described in the literature [13,14,17,18]. Briefly, the first raw moment for $(X)$ gives the expected number of patients who experience an event at time $(t_i)$ within a particular group, and is written as

$$\mu_1' = \mu_{x_i} = E[X = x_i] = \frac{R_{i1} d_i}{R_i}. \quad (3)$$

The finite second central moment is obtained as

$$\mu_2' = E[x_i(x_i - 1)] + E[x_i] = \sum_{x_i=2}^{R_{i1}} \frac{d_i(d_i - 1)\binom{d_i - 2}{x_i - 2}\binom{R_i - d_i}{R_{i1} - x_i}}{\frac{R_i(R_i - 1)}{R_{i1}(R_{i1} - 1)}\binom{R_i - 2}{R_{i1} - 2}} + \frac{R_{i1} d_i}{R_i}. \quad (4)$$

Subtracting the square of the first raw moment from the second central moment gives the variance of $(X)$ at time $(t_i)$, i.e.,

$$\sigma^2_{x_i} = Var[X = x_i] = \frac{d_i R_{i1} R_{i2}(R_i - d_i)}{{R_i}^2(R_i - 1)} = R_{i1}\left[\frac{d_i}{R_i}\right]\left[1 - \left(\frac{d_i}{R_i}\right)\right]\left[\frac{R_i - R_{i1}}{R_i - 1}\right]. \quad (5)$$

2.1.3. Large Sample Properties

Under large sampling conditions, the null distribution for the hypergeometric test may be indirectly approximated by a Gaussian distribution [19]. As $(d_i)$ and $(R_i)$ approach infinity (with a fixed ratio), and assuming that $(R_i - d_i)$ is relatively large, with a fixed finite $(x_i)$, we see that [18]

$$P(X = x_i |\, R_i, R_{i1}, d_i) \sim \left(\frac{d_i!}{(d_i - x_i)!}\right)\left(\frac{R_{i1}!}{x_i!(R_{i1} - x_i)!}\right)\left(\frac{(R_i - d_i)!}{(R_i - d_i - R_{i1} + x_i)!}\right)\left(\frac{(R_i - R_{i1})!}{(R_i)!}\right) \quad (6)$$

$$\sim \left[\frac{\Gamma(R_{i1} + 1)}{\Gamma(x_i + 1)\Gamma(R_{i1} - x_i + 1)}\right]\left[\left(\frac{d_i}{R_i}\right)^{x_i}\left\{\frac{(R_i - d_i)^{R_{i1} - x_i}}{(R_i)^{R_{i1} - x_i}}\right\}\right], \quad (7)$$

$$\sim \left[\left\{\frac{\prod_{j=1}^{x_i}(R_{i1} - j + 2)}{\prod_{j=0}^{x_i - 1}(x_i - j)}\right\} - \left\{\frac{\prod_{j=1}^{x_i - 1}(R_{i1} - j + 1)}{\prod_{j=0}^{x_i - 2}(x_i - 1 - j)}\right\}\right]\left[\left(\frac{d_i}{R_i}\right)^{x_i}\left\{\left(\frac{R_i - d_i}{R_i}\right)^{R_{i1} - x_i}\right\}\right], \quad (8)$$

where the term in the left square brackets of the last two expressions denotes the unordered ways to choose $(x_i)$ from a set of $(R_{i1})$ elements (Pascal's pyramid) [20]. The approximation becomes increasingly better as the ratio terms $\left({R_{i1}}^2/R_i\right)$, $\left({x_i}^2/d_i\right)$, and $\left[(R_{i1} - x_i)^2/(R_i - d_i)\right]$ diminish in size. Applying Stirling's approximation,

$$\gamma! \sim \left(\frac{\gamma}{e}\right)^{\gamma}\sqrt{2\pi\gamma} \quad \text{(with a positive relative error} \leq \frac{1}{12\gamma - 1}\text{)}, \quad (9)$$

we have [21]

$$P(X = x_i |\, R_i, R_{i1}, d_i) \sim \left[\frac{\left(\frac{R_{i1}}{e}\right)^{R_{i1}}\sqrt{2\pi R_{i1}}}{\left(\frac{x_i}{e}\right)^{x_i}\sqrt{2\pi x_i}\left(\frac{R_{i1} - x_i}{e}\right)^{R_{i1} - x_i}\sqrt{2\pi(R_{i1} - x_i)}}\right]\left(\frac{d_i}{R_i}\right)^{x_i}\left(1 - \frac{d_i}{R_i}\right)^{R_{i1} - x_i} \quad (10)$$

$$\sim \left[\frac{R_{i1} d_i}{R_i x_i}\right]^{x_i}\left[\frac{R_{i1}(R_i - d_i)}{R_i(R_{i1} - x_i)}\right]^{R_{i1} - x_i}\sqrt{\frac{R_{i1}}{2\pi x_i(R_{i1} - x_i)}}. \quad (11)$$

Next, we obtain the following identities

$$-log\left(1 + \frac{R_i x_i - R_{i1} d_i}{R_{i1} d_i}\right) = -log\left(\frac{x_i}{R_{i1}\left(\frac{d_i}{R_i}\right)}\right) = log\left(\frac{R_{i1} d_i}{R_i x_i}\right) \quad (12)$$

$$\Rightarrow log\left(\frac{R_{i1}(R_i - d_i)}{R_i(R_{i1} - x_i)}\right) = -log\left(1 - \frac{x_i - R_{i1}\left(\frac{d_i}{R_i}\right)}{R_{i1}\left(1 - \frac{d_i}{R_i}\right)}\right) = -log\left(1 - \frac{R_i x_i - R_{i1} d_i}{R_{i1}(R_i - d_i)}\right). \quad (13)$$

Noting that

$$-\sum_{j=1}^{\infty} \frac{(-1)^j a^{-j}}{j} = log(1+a) \text{ for } |a| > 1 \tag{14}$$

and combining terms, with $O\left(\frac{(x_i - R_{i1}(d_i/R_i))^3}{{R_{i1}}^2}\right)$ dominating over $O\left(\frac{(x_i - R_{i1}(d_i/R_i))^3}{{R_{i1}}^3}\right)$, it follows that

$$\left[\frac{R_{i1}d_i}{R_i x_i}\right]^{x_i}\left[\frac{R_{i1}(R_i - d_i)}{R_i(R_{i1} - x_i)}\right]^{R_{i1}-x_i} = exp\left(-\frac{1}{2}\left[\frac{x_i - R_{i1}\left(\frac{d_i}{R_i}\right)}{\sqrt{R_{i1}\left(\frac{d_i}{R_i}\right)\left(1-\frac{d_i}{R_i}\right)}}\right]^2\right). \tag{15}$$

Continuing, we assume that

$$x_i - R_{i1}\left(\frac{d_i}{R_i}\right) \approx \sqrt{R_{i1}}. \tag{16}$$

When $\left(\frac{d_i}{R_i}\right)$ is neither close to 0 nor 1 and both $\left[R_{i1}\left(\frac{d_i}{R_i}\right)\right]$ and $\left[R_{i1}\left(1-\left(\frac{d_i}{R_i}\right)\right)\right]$ are large, the application of L'Hôpital's rule shows that

$$\lim_{R_{i1}\to\infty}\left[\frac{R_{i1} - R_{i1}\left(\frac{d_i}{R_i}\right)}{x_i}\right] = \lim_{R_{i1}\to\infty}\left[\frac{1-\left(\frac{d_i}{R_i}\right)}{\frac{1}{2\sqrt{R_{i1}}} + \left(\frac{d_i}{R_i}\right)}\right] \to 1. \tag{17}$$

Therefore,

$$R_{i1} - x_i \approx R_{i1}\left(1-\left(\frac{d_i}{R_i}\right)\right) - \sqrt{R_{i1}}\left[\frac{R_{i1}\left(1-\left(\frac{d_i}{R_i}\right)\right)}{x_i}\right] \tag{18}$$

$$\implies x_i(R_{i1} - x_i) \approx \left(x_i - \sqrt{R_{i1}}\right)\left[R_{i1}\left(1-\left(\frac{d_i}{R_i}\right)\right)\right] \approx (R_{i1})^2\left(\frac{d_i}{R_i}\right)\left(1-\left(\frac{d_i}{R_i}\right)\right). \tag{19}$$

Substituting accordingly and combining terms gives

$$P(X = x_i |\, R_i, R_{i1}, d_i) \;\sim\; \sqrt{\frac{1}{2\pi R_{i1}\left(\frac{d_i}{R_i}\right)\left(1-\left(\frac{d_i}{R_i}\right)\right)}} exp\left(-\frac{1}{2}\left[\frac{x_i - R_{i1}\left(\frac{d_i}{R_i}\right)}{\sqrt{R_{i1}\left(\frac{d_i}{R_i}\right)\left(1-\frac{d_i}{R_i}\right)}}\right]^2\right) \;\sim\; N\left(\mu_{x_i}, \sigma_{x_i}^2\right), \tag{20}$$

where $\mu_{x_i}$=$\left(\frac{R_{i1}d_i}{R_i}\right)$ and $\sigma_{x_i}^2 = R_{i1}\left(\frac{d_i}{R_i}\right)\left(1-\frac{d_i}{R_i}\right)$. Therefore, the discrete probability elements for each $(X)$ at time $(t_i)$ shrink infinitesimally to yield a symmetrical continuous density centered at $(\mu_{x_i})$ with asymptotic points of inflection at $\left(\mu_{x_i} \pm \sigma_{x_i}^2\right)$. A simple transformation gives

$$\xi_i = \frac{(x_i - \mu_{x_i})}{\sigma_{x_i}} \;\sim\; N(0,\, 1). \tag{21}$$

Noting that $\lim_{R_i\to\infty}\left[{(R_{i1})^3}/{(R_i)^2}\right] \to 0$, the $\left[\frac{R_i - R_{i1}}{R_i - 1}\right]$ term in variance for the hypergeometric distribution asymptotically approaches unity, and, as expected, the corresponding variance for the Gaussian distribution becomes $\left[R_{i1}\left[\frac{d_i}{R_i}\right]\left[1-\left(\frac{d_i}{R_i}\right)\right]\right]$. Lastly, we mention that a more direct proof yielding the normal distribution can be obtained by rewriting

the binomial coefficients in the hypergeometric distribution using de-Moivre Laplace's asymptotic formula and simplifying [22].

#### 2.1.4. Useful Approximations, Bounds, and Recursive Formulas

When $(R_i > 50), (d_i \leq R_{i1})$, and $\left(\frac{2d_i-\eta}{2R_i-R_{i1}+1} \leq 1\right)$, a reasonable approximation for the sum of hypergeometric terms, in terms of the Bernoulli distribution, is given as [14]

$$P(X \leq r|\ R_i, d_i, R_{i1}) = \sum_{x_i=max[0,\ R_{i1}-(R_i-d_i)]}^{r} \frac{\binom{d_i}{x_i}\binom{R_i-d_i}{R_{i1}-x_i}}{\binom{R_i}{R_{i1}}} \tag{22}$$

$$= 1 - \sum_{x_i=r+1}^{R_{i1}} \left[\left[\frac{\prod_{j=1}^{x_i}(R_{i1}-j+2)}{\prod_{j=0}^{x_i-1}(x_i-j)}\right] - \left[\frac{\prod_{j=1}^{x_i-1}(R_{i1}-j+1)}{\prod_{j=0}^{x_i-2}(x_i-1-j)}\right]\right]\left[\left(\frac{d_i}{R_i}\right)^{x_i}\left[\frac{(R_i-d_i)^{R_{i1}-x_i}}{(R_i)^{R_{i1}-x_i}}\right]\right] \tag{23}$$

$$\sim 1 - \sum_{x_i=r+1}^{R_{i1}} \left[\left[\frac{\prod_{j=1}^{x_i}(R_{i1}-j+2)}{\prod_{j=0}^{x_i-1}(x_i-j)}\right]\left[\frac{\prod_{j=1}^{x_i-1}(R_{i1}-j+1)}{\prod_{j=0}^{x_i-2}(x_i-1-j)}\right]\right]\left[\left(\frac{2d_i-r}{2R_i-R_{i1}+1}\right)^{x_i}\left(\frac{2(R_i-d_i)+(1+r)-R_{i1}}{2R_i-R_{i1}+1}\right)^{R_{i1}-x_i}\right]. \tag{24}$$

A lower and upper bound for the hypergeometric density, as a function of the Bernoulli distribution is written as [18]

$$\begin{aligned} exp\left[-\tfrac{1}{2}\tfrac{x_i(x_i-1)}{2(d_i-x_i)} - \tfrac{1}{2}\tfrac{(R_{i1}-x_i)(R_{i1}-x_i-1)}{(R_i-d_i-R_{i1}+x_i)}\right]\binom{R_{i1}}{x_i}\left(\tfrac{R_{i1}d_i}{R_i}\right)^{x_i}\left(\tfrac{R_i-R_{i1}d_i}{R_i}\right)^{R_{i1}-x_i} \\ \leq \frac{\binom{d_i}{x_i}\binom{R_i-d_i}{R_{i1}-x_i}}{\binom{R_i}{R_{i1}}} \leq exp\left[\tfrac{R_{i1}(R_{i1}-1)}{2(R_i-R_{i1})}\right]\binom{R_{i1}}{x_i}\left(\tfrac{R_{i1}d_i}{R_i}\right)^{x_i}\left(\tfrac{R_i-R_{i1}d_i}{R_i}\right)^{R_{i1}-x_i}. \end{aligned} \tag{25}$$

This readily follows from the inequality

$$exp\left[\frac{-R_{i1}(R_{i1}-1)}{2(R_i-R_{i1})}\right] \leq \frac{(R_i)_{R_{i1}}}{(R_i)^{R_{i1}}} \leq \left(\frac{R_i-R_{i1}d_i}{R_i}\right)^{R_{i1}}\sum_{x_i=0}^{R_{i1}}\binom{R_{i1}}{x_i}\left(\frac{R_{i1}d_i}{R_i-R_{i1}d_i}\right)^{x_i}. \tag{26}$$

In many cases, determining hypergeometric probabilities can be challenging. A convenient recursive equation is easily derived as

$$\begin{aligned} P(X=x_i+1|\ R_i, R_{i1}, d_i) =& P(X=x_i|\ R_i, R_{i1}, d_i)\left[\tfrac{d_i!}{(x_i+1)!(d_i-x_i-1)!}\right]\left[\tfrac{x_i!((d_i-x_i)!)}{d_i!}\right] \\ &\times\left[\tfrac{(R_i-d_i)!}{(R_{i1}-x_i-1)!(R_i-d_i-R_{i1}+x_i+1)!}\right]\left[\tfrac{(R_{i1}-x_i)!(R_i-d_i-R_{i1}+x_i)!}{(R_i-d_i)!}\right] \end{aligned} \tag{27}$$

$$P(X=x_i|\ R_i, R_{i1}, d_i)\left[\frac{(R_{i1}-x_i)(d_i-x_i)}{(x_i+1)(R_i-d_i-R_{i1}+x_i+1)}\right]. \tag{28}$$

Rearranging, we see that

$$P(X=x_i-1|\ R_i, R_{i1}, d_i) = P(X=x_i|\ R_i, d_i, R_{i1})\left(\tfrac{x_i}{d_i-x_i+1}\right)\left(\tfrac{R_i-d_i-R_{i1}+x_i}{R_{i1}-x_i+1}\right). \tag{29}$$

### 2.2. Weighted Log-Rank Test

Consider $(m)$ separate event time points $(t_1 < t_2 < t_i < \cdots < t_m)$ and let $(w_i)$ denote a non-disjoint, positive weight function that is appropriately bounded (detectable, non-zero measure) for each $(i)$ value. The linear combination $(\sum_i^m w_i\xi_i)$ yields the weighted LRT, which defaults to the standard LRT when the weight function is equal to unity for each time point [23,24]. Because the moment generating function (MGF) for this linear

combination is equal to the MGF of a normal distribution with mean = $(\sum_i^m w_i\mu_{x_i})$ and variance = $(\sum_i^m {w_i}^2\sigma_{x_i}^2)$, i.e.,

$$MGF_X(s) = e^{[s(\sum_{i=1}^m w_i\mu_{x_i})+\frac{s^2}{2}(\sum_{i=1}^m {w_i}^2\sigma_{x_i}^2)]}, \quad (30)$$

it holds that

$$\sum_{i=1}^m w_i\xi_i \sim N\left(\sum_{i=1}^m w_i\mu_{x_i}, \sum_{i=1}^m {w_i}^2\sigma_{x_i}^2\right), \quad (31)$$

since no distinct probability distributions can have the same moment generating function. Thus, under large sampling conditions, the summation of ( $w_i\xi_i$) over $(m)$ time points has an approximate standard normal distribution, i.e., $N(0, 1)$ or, equivalently, by taking the square, a chi-square distribution with one degree of freedom.

Rewriting the weighted LRT as

$$\xi_{w_i} = \sum_{i=1}^m \frac{\left[(w_i)\left(d_{i1} - \frac{R_{i1}d_i}{R_i}\right)\right]^2}{\sum_{i=1}^m {w_i}^2\sigma_{x_i}^2} = \sum_{i=1}^m \frac{[w_i(\,O_i - E_i)]^2}{\sum_{i=1}^m Var[w_i(O_i - E_i)]}, \quad (32)$$

where $(O_i - E_i)$ denotes the deviation of the observed values $(d_{i1})$ from their expected values, we see that the numerator of $(\xi_{w_i})$ corresponds to the weighted sum of conditionally independent and uncorrelated hypergeometric (asymptotically normal) random variables, with each term having a mean of zero, under the null hypothesis of no treatment effect (i.e., $E[w_i d_{i1_i}] = \frac{w_i R_{i1} d_i}{R_i}$) [10]. Since the event times are conditionally independent of one another and are functionally predictable (i.e., $\xi_{w_i}$ is not contingent on outcomes that occur at or beyond $t_i$) [25], the variance of the numerator is simply equal to the sum of the variances for the individual $[w_i(O_i - E_i)]$ terms [15]. Specifically,

$$Var[w_i(O_i - E_i)] = {w_i}^2[Var(O_i) + Var(E_i) - 2Cov(O_i,\ E_i)] = {w_i}^2 Var(O_i), \quad (33)$$

as both the variance of $(E_i)$ and the $Cov(O_i,\ E_i)$ are equal to zero. Of further note, $(\xi_{w_i})$ remains the same if $(w_i)$ is multiplied or divided by a scaler constant [26,27].

Applying the conditional central limit theorem (assuming the exchangeability of elements and Lundeberg's sufficiency conditions for martingales—finite variance, tightness, and uniform integrability), it follows that $(\xi_{w_i})$ is asymptotically consistent and weakly convergent in distribution to a chi-square distribution with 1 degree of freedom, even when the individual terms are not necessarily identically distributed [28–33]. Thus, the conditional central limit theorem aligns with the abovementioned MGF approach for defining the large sample distribution of $(\xi_{w_i})$ but with less stringent conditions that are better suited for real-world applications [34]. Nonetheless, the small-sample behavior in both scenarios may be difficult to anticipate in practice, especially for highly censored and sparse tailed data [35].

### 2.3. Selection of Weights

Various choices for $(w_i)$ have been proposed in the literature. A popular selection is to set $(w_i)$ equal to 1, which gives the standard Mantel–Haenszel LRT without continuity correction [23]. While this option is fairly robust for detecting survival curve differences, especially in the case of proportional hazards, there is no universal consensus regarding the best weight or combination of weights to use when the hazards (for the two groups under comparison) are not constant over time, as is the case for late separating survival curves. One flexible option is the two-parameter Fleming–Harrington (FH) weight, with $(w_i)$ defined as

$$G(\rho,\ \gamma) = \left\{\widetilde{S}(t-)^{\rho}\right\}\left\{1 - \widetilde{S}(t-)^{\gamma}\right\}, \quad (34)$$

where $\widetilde{S}(t-)$ is the left-continuous product-limit estimate, and $(\rho \geq 0,\ \lambda \geq 0)$ [29]. Here, $G(\rho = 0, \gamma = 0)$, $G(\rho > 0, \gamma = 0)$, $G(\rho > 0, \gamma > 0)$, and $G(\rho = 0, \gamma > 0)$ purportedly corresponds to "evenly distributed", "early", "mid", and "late" treatment effects, with $G(\rho = 0, \gamma = 0)$ denoting the standard LRT and $G(\rho = 1, \gamma = 0)$ denoting the Prentice–Wilcoxon statistic. Barring prior knowledge, the selection of $(\rho)$ and $(\gamma)$ is largely arbitrary. Arguably, certain weights may lack clinical relevance, focusing only on a specific portion of a survival curve with low event rates or diminishing treatment effects.

A compromise entails taking the maximum of the standardized statistics for a preset combination of FH-LRT values for $(\rho)$ and $(\gamma)$. Dividing the difference vector by the corresponding square root of Fisher's information matrix (a non-singular, uniformly minimum variance unbiased estimator), the resultant statistic asymptotically assumes a multivariate Gaussian distribution [36]. Known as the "Max-Combo" method, the test accommodates various treatment effects by selectively up- or down-weighting the log-rank statistics over time [37]. In general, combination approaches are more powerful than the standard LRT under a range of nonproportional hazard conditions [38,39]. The critical value $(c_\alpha)$ for a $(k)$-component Max-Combo test $\left(Z^k_{Max}\right)$ is defined such that

$$P\left[max\left\{Z_1,\ Z_1,\ \ldots\ Z_k,\ \right\} \geq c_\alpha\right] = \alpha. \tag{35}$$

Commonly used combinations include

$$Z^3_{Max} = max\{G(0,0), G(1,0), G(0,1)\}; \tag{36}$$

$$Z^4_{Max} = max\{G(0,0), G(0,1), G(1,1), G(1,0)\}; \tag{37}$$

$$Z^4_{Max} = max\{G(0,0), G(2,0), G(0,2), G(2,2)\}, \tag{38}$$

with the first abovementioned $Z^4_{Max}$ traditionally being designated as the default set of weights. The Max-Combo test has been shown to perform well in many applied examples with non-proportional hazards [40]. However, under moderate to heavy censoring and noting the potentially high correlation among weighted LRTs, the family of combination procedures (including the Max-combo test) may not be more versatile than individual component LRT tests [8]. The extension to a group sequential analysis allows the Max-Combo procedure to accommodate multiple time point decisions, with the test statistic assuming a joint normal distribution under the null hypothesis (per the application of Slutsky's theorem) [41–44].

### 2.4. Inverse Log-Rank Test

A key constraint of the Max-Combo test in practical applications is that the null hypothesis can be rejected in favor of both the experimental and reference arms for an identical set of observations [45]. That is, when survival curves cross and one wishes to test the superiority of Treatment A, it is possible for the Max-Combo method to reject the null hypothesis in favor of Treatment A; while in contrast, if the objective is to test the superiority of Treatment B, then the Max-Combo method could conceivably yield the opposite conclusion given the same data (i.e., reject the null hypothesis in favor of Treatment B). Alternatively, the iLRT presents a single-weight LRT for analyzing non-proportional hazard survival curves [46].

Based on a smoothed, non-negative function of sample values that converges in probability to its true state, the inversely weighted logarithm of the combined number of patients at risk at each of $(m)$ study time points is given by

$$w_i = \frac{\log(R_i)}{R_i} \quad (i = 1\ to\ m). \tag{39}$$

The iLRT is defined as

$$\zeta_{w_i} = \sum_{i=1}^{m} \frac{\left[(w_i)\left(d_{i1} - \frac{R_{i1}d_i}{R_i}\right)\right]^2}{\sum_{i=1}^{m}(w_i)^2\sigma_{x_i}^2} = \sum_{i=1}^{m} \frac{\left[\left(\frac{log(R_i)}{R_i}\right)\left(d_{i1} - \frac{R_{i1}d_i}{R_i}\right)\right]^2}{\sum_{i=1}^{m}\left(\frac{log(R_i)}{R_i}\right)^2\sigma_{x_i}^2}, \quad (40)$$

with the $p$-value (two-sided test) estimated as

$$P\left(\mathcal{X}_1^2 > \zeta_{w_i}\right) = \int_{x=\zeta_{w_i}}^{\infty} \frac{e^{-x/2}}{\sqrt{2\pi x}} dx. \quad (41)$$

As $(\pi)$ in the denominator of the last equation is equal to $\Gamma(1/2)$, the integrand corresponds to the probability density function of a chi-square distribution with 1 degree of freedom. Being a score that is statistic, which can be alternatively expressed as a discrete-time, partial likelihood function, $(\zeta_{w_i})$ easily accommodates censored data [47,48].

### 2.5. Split-Range Test

Consider the special case of a 2-arm, randomized clinical trial where all of the patients in the comparison arm (Group 2) achieve the event of interest by a certain time, while some of the patients in the test arm (Group 1) have survival times beyond this time point. A $p$-value for testing the null hypothesis $(H_0)$ of no survival differences between the groups may be computed using the split-range test (SRT) [49]. In this non-parametric method, designate the number of patients in Group 1 as $(n_1)$ and the number in Group 2 as $(n_2 = N - n_1)$, with $(N)$ denoting the total sample size. This is equivalent to the Fermi–Dirac "ball and cell" model, where $(n_2)$ balls are randomly dropped into $(N)$ cells (corresponding to ranked survival times), allowing one ball per cell. Numbering the cells from 1 to $(N)$, the range $(R)$ is defined as the number of the highest occupied cell minus the lowest occupied cell. The value for the range must be a number from $(n_2 - 1)$ to $(N - 1)$. To test $(H_0)$ with a Type I error rate of $(\alpha)$ for falsely rejecting the null hypothesis, find the integer $(\varphi)$ satisfying

$$\sum_{r=n_2-1}^{\varphi} P(R = r) = \alpha \quad (42)$$

and reject $(H_0)$ if the observed value of $(R)$ does not exceed $(\alpha)$. When censored values occur in Group 1 before the last event occurs in the Group 2, then $(\alpha)$ denotes an upper bound. That is, some of the true survival times for these censored values may be longer than all the elements in Group 1. By decreasing the range, this results in a smaller $p$-value.

Analogously, the split-range test can be applied in reverse by randomly dropping the $(n_1)$ balls into the $(N)$ cells. Again, the range is defined as the number of the highest occupied cell minus the lowest occupied cell. A non-directional test is obtained by simultaneously considering both cases and multiplying $(\alpha)$ by two to adjust for multiplicity.

### 2.6. Computational Details

$p$-Values for the weighted Fleming and Harrington LRT were computed using the "Test=FH" option in the strata statement of the LIFETEST procedure in SAS v.9.4 software (Cary, NC, USA), while $p$-values for the Max-Combo procedure were obtained iteratively [50]. The SAS code for performing the iLRT is provided in the Appendix A. In most cases, the computational run time for the iLRT is approximately 4-fold (or more) faster than the default 4-component Max-Combo test.

$p$-Values $\leq 0.05$ were deemed to be statistically significant. Unless otherwise indicated, computed values were presented to two significant digits using the Goldilocks (Efron–Whittemore) rounding method, rather than a fixed number of decimal places [51].

## 3. Examples

Four examples are presented in this section comparing the results of the Prentice–Wilcoxon, standard Mantel (unweighted), combination Max-Combo (default four-component), and inverse log-rank tests. The combinatoric SRT is presented as a non-LRT comparison in the fourth example. Kaplan–Meier (product-limit) plots are provided for each example (see Figure 1). Summary computational results of the iLRT for the four examples are shown in Table 2.

**Table 2.** Summary computations for the inverse log-rank test (Examples 1–4).

| Ex. | Group 1 | | Group 2 | | Numerator $\left[\sum_{i=1}^{m} w_i(d_{i1}-\mu_{x_i})\right]^2$ | Denominator $\sum_{i=1}^{m}(w_i)^2\sigma_{x_i}^2$ | $\chi_1^2$ | $p$ * |
|---|---|---|---|---|---|---|---|---|
| | Event | Censored | Event | Censored | | | | |
| 1 | 64 | 11 | 82 | 0 | 1.6 | 0.33 | 4.8 | 0.029 |
| 2 | 57 | 2 | 63 | 0 | 0.017 | 0.37 | 0.046 | 0.37 |
| 3 | 16 | 9 | 5 | 19 | 0.46 | 0.047 | 9.9 | 0.0017 |
| 4 | 88 | 12 | 100 | 0 | 2.0 | 0.19 | 11 | 0.0011 |

* $p$-Values computed using non-rounded values. Ex. = Example. $m$ = # of time points; $w_i = \frac{\log(R_i)}{R_i}$; $\mu_{x_i} = \frac{R_{i1}d_{i1}}{R_i}$; $\sigma_{x_i}^2 = R_{i1}\left[\frac{d_i}{R_i}\right]\left[1-\left(\frac{d_i}{R_i}\right)\right]\left[\frac{R_i-R_{i1}}{R_i-1}\right]$; $\chi_1^2$ = Numerator/Denominator.

### 3.1. Example 1

In this non-randomized cohort of $n$ = 157 emulated patients with metastatic (stage IV), non-squamous cell lung cancer (NSCLC), who failed to respond to conventional chemotherapy, 75 opted to receive an experimental immune therapy compound (Group 1) versus 82 who were provided hospice care (Group 2) [46]. Among the 75 patients in the first group, 11 had censored outcomes, while all of the patients in Group 2 experienced an event (Table 2). Soon after the second month, a noticeable late survival advantage materialized for the experimental group, while those in the hospice group continued to decline (see Kaplan–Meier plot for Example 1). Notably, the Kaplan–Meier curves otherwise crisscrossed for the first two months before diverging. The median survival time for Group 1 was slightly higher than Group 2 (0.69 versus 0.65 months). Only the iLRT yielded a statistically significant survival group difference ($p$ = 0.029). Although the default Max-Combo failed to achieve statistical significance ($p$ = 0.071), several individual FH-LRT values for ($\rho$) and ($\gamma$) had correspondingly lower $p$-values than the iLRT, with a minimum being observed for ($\rho$ = 0, $\gamma$ = 5; $p$ = 0.015) (Table 3). That is, the power of the Max-Combo test in a specific scenario may not exceed their component FH test statistics [52].

**Table 3.** Individual FH-LRT $p$-values for G($\rho$, $\gamma$).

| G(0, γ) | *p*-Value | G(ρ,5) | *p*-Value |
|---|---|---|---|
| (0,0) | 0.27 | (0,5) | 0.015 |
| (0,1) | 0.033 | (1,5) | 0.44 |
| (0,5) | 0.015 | (5,5) | 0.35 |
| (0,10) | 0.021 | (10,5) | 0.18 |
| (0,15) | 0.036 | (15,5) | 0.47 |
| (0,20) | 0.053 | (20,5) | 0.85 |
| (0,25) | 0.069 | (25,5) | 0.42 |

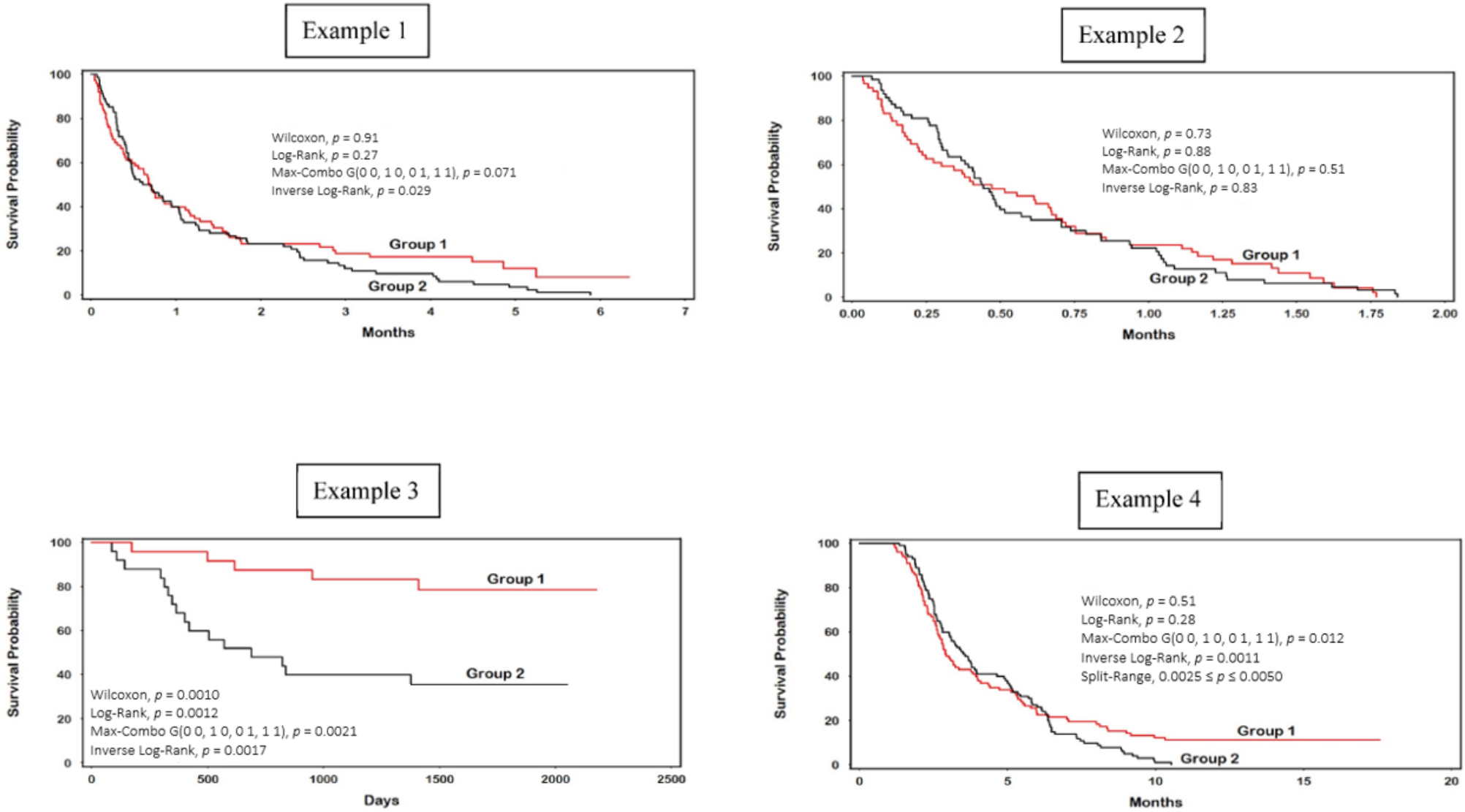


**Figure 1.** Kaplan–Meier curves corresponding to Examples 1–4.

### 3.2. Example 2

The objective of Example 2 is to demonstrate the non-significant difference between the two treatment arms in Example 1, prior to their point of separation. As expected, upon deleting observations occurring after 1.9 months, none of the LRTs in this example had statistically significant $p$-values. The highest value $p$-value corresponded to the iLRT ($p = 0.83$), followed by the default Max-Combo test ($p = 0.51$).

### 3.3. Example 3

An important characteristic of an omnibus LRT is the ability to accommodate late separating survival curves, while also having power to detect significant differences occurring from the beginning of a study. Example 3 elaborates on the comparative analysis of two cancer therapies, historically presented by Brown and Hollander [53]. Referring to the Kaplan–Meier plots for this example, we see that the treatment curves are relatively parallel, suggesting proportional hazards over time. Both the standard Mantel LRT ($p = 0.0012$) and Prentice–Wilcoxon LRT ($p = 0.0010$) are statistically significant, while the iLRT ($p = 0.0017$) and the default Max-Combo test ($p = 0.0021$) yield comparable levels of statistical significance, though to a slightly lesser degree.

### 3.4. Example 4

Example 4 illustrates a special case of late separating survival curves, as originally presented by the author [49]. In this analysis, all the patients in the comparison arm (Group 2) experience the event of interest, while 11 of the patients in the experimental treatment arm (Group 1) have survival times greater than the last event in Group 2 at 9.5 years. Accordingly, the SRT is applicable in this example and yields a $p$-value of between 0.0025 and 0.0050, as there is one censored value at 3.0 years that occurs in Group 1 before the last event in Group 2. While all of the values in Group 1 beyond the completion of Group 2 are censored, an equivalent $p$-value would have been obtained for this degenerate case, even if one or more of these censored values were events (which is the case for LRTs in general).

In this example, the $p$-value obtained for the SRT is comparatively close to the iLRT ($p = 0.0011$) and the default (four-component) Max-Combo procedure ($p = 0.012$), with the iLRT yielding the more statistically significant value. The cumulative frequency for the split-range test given $n = 100$ and $N = 200$ is provided in Table 4.

**Table 4.** Cumulative frequency for the split-range test (n = 100, N = 200).

| $r$ | $P(R{=}r)$ | $P(R{\leq}r)$ | $r$ | $P(R{=}r)$ | $P(R{\leq}r)$ |
|---|---|---|---|---|---|
| 184 | 0.00008 | 0.00016 | 192 | 0.01498 | 0.03243 |
| 185 | 0.00016 | 0.00032 | 193 | 0.02677 | 0.05920 |
| 186 | 0.00032 | 0.00064 | 194 | 0.04662 | 0.10582 |
| 187 | 0.00063 | 0.00127 | 195 | 0.07851 | 0.18434 |
| 188 | 0.00122 | 0.00249 | 196 | 0.12627 | 0.31060 |
| 189 | 0.00234 | 0.00483 | 197 | 0.18940 | 0.50000 |
| 190 | 0.00441 | 0.00924 | 198 | 0.25126 | 0.75126 |
| 191 | 0.00821 | 0.01745 | 199 | 0.24874 | 1.0000 |

### 3.5. Comparison with the Cox Regression Model

In Examples 1, 2, and 4, which depict non-proportional hazards, the corresponding hazard ratios (HRs) and significance levels (estimated by a Cox regression model) were 1.2 ($p = 0.27$), 1.0 ($p = 0.88$), and 1.2 ($p = 0.28$), respectively. In contrast, the hazards for the two survival curves shown in Example 3 were relatively constant over time (HR = 0.22) and

manifested a $p$-value of 0.0030, being slightly less significant but comparable to the iLRT ($p$ = 0.0017) and default Max-Combo procedure ($p$ = 0.0021).

## 4. Sample Size and Power

### 4.1. Sample Size and Power Methodology

To compute the sample size and power for a planned trial (i.e., how frequently a test will detect the falsehood of an underlying hypotheses when it is wrong), we note that [54]

$$\Psi = Z_{\alpha/2} + Z_{\beta}, \tag{43}$$

where $(\Psi)$ is the standardized test statistics for the iLRT, and $Z_{\alpha}$ denotes the $100(1-\alpha)$ percentile of a standard normal distribution and proceed in a manner comparable to Garès and colleagues [55]. Specifying the desired power as $(1-\beta)$ for an $\alpha$-level (two-sided) test of significance, the respective sample size for Group 1 $\left(N_{\boldsymbol{T}otal} = 2 \times N_{Group\ 1}\right)$ is given as

$$N_1 = \left[\frac{\sigma(Z_{\alpha/2} + Z_{\beta})}{\sum_{i=1}^{m}\left(\frac{log(R_i)}{R_i}\right)\left(d_{i1} - \frac{R_{i1}d_i}{R_i}\right)}\right]^2, \tag{44}$$

where

$$\sigma = \sqrt{R_{11}\sum_{i=1}^{m}\left(\frac{log(R_i)}{R_i}\right)^2 \sigma_{x_i}^2}. \tag{45}$$

Rearranging the formula for sample size, we see that

$$\text{Power} = (1-\beta) = \left\{0.5 + 0.5erf\left[\left\{\left[\frac{\sqrt{N_1}}{\sigma}\sum_{i=1}^{m}\left(\frac{log\left(R_i\right)}{R_i}\right)\left(d_{i1} - \frac{R_{i1}d_i}{R_i}\right)\right] - Z_{\alpha/2}\right\}/\sqrt{2}\right]\right\}, \tag{46}$$

where

$$erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt. \tag{47}$$

### 4.2. Sample Size and Power Example

In Example 1, the results of a non-randomized cohort were presented where a new experimental compound was compared with hospice care for late stage, refractory lung cancer. Based on the promising findings from this study, a pharmaceutical company would like to conduct a Phase-3 clinical trial randomizing an equal number of patients to the two treatment groups.

Specifically, the company wishes to reject the null hypothesis of equivalent survival times between the two arms of the planned study with a probability of 90% (given that the survival curves are truly different), and a Type I (two-sided) error rate of 5%. Plugging in the numbers from the first row of Table 1, we see that

$$N_1 = \left[\frac{\sqrt{(64+11)\ (0.33)}(1.96+1.3)}{\sqrt{1.6}}\right]^2 \cong 164. \tag{48}$$

Upon being informed of the sample size, management decided that the cost to conduct the trial would be too high. Instead, they suggested a trial of no more than 144 patients per arm and asked the statistician to determine the corresponding statistical power, computed as

$$\text{Power} = (1-\beta) = \left\{0.5 + 0.5erf\left[\left\{\left[\frac{(12)(1.26)}{4.97}\right] - 1.96\right\}/1.41\right]\right\} \cong 73\%. \tag{49}$$

## 5. Discussion

### 5.1. Overview

The choice of weights for an LRT is arbitrary and largely predicated on the efficiency to detect treatment differences [56]. Under the null hypothesis, optimal "pre-specified" weights are a function of the total number of participants at risk at the time of a respective event and are estimated from the data [57]. While weighted rank tests are valid under unequal censoring, the asymptotic relative efficiency of the test statistic depends on the censoring distribution. A weighted LRT should be reasonably robust to unequal right-censoring, as permutation tests may fail to provide suitable approximations [58]. In such cases, the permutation computed variance may underestimate the true variance when censoring is unequal [59]. Additionally, the analysis of arbitrarily interval-censored survival data requires special techniques beyond that discussed here [60,61].

The optimal weight or combination of weights for an LRT has a defined power advantage, contingent upon advanced knowledge of when the survival curve separation will occur (e.g., early, mid, or late). Thus, the ideal selection depends on the data, knowledge of which may not be feasible before the completion of a study. While pilot data or results from comparable studies can be helpful in the decision-making process, there is no guarantee that a planned study will behave similarly. While several researchers have proposed adaptively choosing weights as a function of the data [38,47,62], the properties of such tests may be challenging to predict and may have less power when compared with the traditional unweighted LRT with proportional hazards [25].

The iLRT is nearly as powerful as the standard LRT under proportional hazards. Yet, the iLRT is more sensitive to time-dependent, non-proportional hazards observed for differential or single arm delayed treatment effects. When an investigator is uncertain in advance about the shape of the survival curves, it is not apposite to select an LRT after the data have been collected as the analytic method should be clearly specified in the protocol prior to the initiation of a study. One option is to select a combination of FH weights in the form of the Max-Combo test. While this procedure performs reasonably well, again as previously noted, it is possible to reject the null hypothesis both in favor and against a particular treatment for the same data [45]. Combination tests also may have diminished power, albeit marginal, to detect treatment differences, resulting from the implicit multiplicity correction required by the procedure. As a single weight method, the iLRT does not require adjustment for multiple testing and provides a flexible and non-subjective means for analyzing both continuing and late separating survival curves. However, if the investigator is certain of the shape of the survival curves in advance, then an appropriately parametrized FH-LRT may present the optimal choice for the planned analysis.

### 5.2. Efficiency

The chi-square statistic $(\zeta_{w_i})$, $\left[\widehat{w_i d_i / R_i}\right]$ is the minimum, best asymptotic normal (BAN) estimator for $[E_i = (w_i R_{i1} d_i / R_i)]$, providing that it is a consistent estimate of the latter and asymptotically normal under large sample conditions (with properties akin to the maximum likelihood estimator and Fisher's information loss, albeit based on cell frequencies vs. original observations) [63–65]. Among all such asymptotically normal estimates within a multinomial framework, none have a smaller variance [66]. As such, $(\zeta_{w_i})$ belongs to a class of tests which are unbiased and equivalent in limit to Neyman's $\lambda$-test [54,67]. While tests within this family have comparable or more stringent power against Pitman alternatives (i.e., asymptotic relative efficiency), there is no guarantee that the statistic converges to a normal distribution at a reasonably fast rate, especially when observations are sparse toward the extreme right tail, with manifest censoring [68–71]. For Type II right-censored data with a presumed number of events, the total time of the trial is unknown until the last event occurs (versus trials with a fixed time of termination) [72]. Nonetheless, both types of censoring may lead to unreliable inferences and are challenging to model if censoring is sporadic, non-stationary, or a differential censoring mechanism exists between the two arms of a trial [73]. The misspecification of weights with respect

to censoring or premature withdrawals can have undesirable and difficult to predict consequences on test efficiency and power, especially in the presence of incomplete data.

### 5.3. Power

#### 5.3.1. Lakatos–Cantor Method for Computing Power

In practice, an alternative method for computing the power of weighted LRTs exists that only requires specifying the survival probabilities at designated times for the two arms being compared. This method (based on a seminal paper by Lakatos in 1988 and later simplified by Cantor for practical application) involves partitioning the study period into a set number of subintervals [74,75]. The survival distribution for each treatment group is approximated by a piecewise linear curve, with the respective hazard at each time point estimated by linear interpolation. A Markov chain process is used to model state transitions of events across time. When both the sample size and corresponding number of subintervals are reasonably large, the power obtained by this method will tend toward that described in Section 4 [76].

The advantage of the piecewise linear approach for determining power is that one can visually estimate the required survival probabilities from published Kaplan–Meier curves or, alternatively for smaller sample sizes, by the Nelson–Aalen method [77]. Furthermore, computer packages for implementing the Lakatos model, allowing for user-provided LRT weights, are readily available [78,79]. The main limitation of this method lies in partitioning the study period into subintervals (i.e., discretizing continuous data into bins), particularly when the number of subintervals is small. In this case, the resulting values within each subinterval can vary depending on how the boundaries for the subintervals are chosen and potentially bias the analysis (i.e., "Mendel effect") [80,81]. Implementing a prescribed algorithm to choose the interval widths alleviates this concern to some degree. However, there is no consensus on the optimal vs. practical approach for binning, with some historic and hitherto commonly used procedures lacking statistical consistency [82–86].

#### 5.3.2. Interim Power and Sample Size Re-Estimation

While event level information often is not available during the planning stage of a clinical trial, investigators typically will have access to published Kaplan–Meier survival plots from previous studies [78]. A stop-gap measure, pending the availability of more precise information, involves initially estimating power using the Lakatos–Cantor method and then re-estimating the power and sample size at an interim point, implementing the iLRT method described in Section 4. Providing that the investigator and other members of the study team remain blinded, there is no need to apply a $p$-value penalty for each interim look at the data.

A first interim analysis typically is conducted after more than half of the planned events in the trial have been observed, with less than ~6% (or a predetermined percentage) of participants being lost to follow-up or early censoring. In some cases, if allowed by the protocol and appropriately penalized, the unblinded "data monitoring committee statistician" may recommend a second sample size re-estimation after 75% of the planned events have occurred since sample sizes may have to be adjusted depending upon the point of late separation for the survival curves. Of note, "writing back" the time of censoring to the time of an earlier administrative event can lead to an artifactual late separation of survival curves or unintended differential bias [87].

### 5.4. Limitations

#### 5.4.1. Potential Sources of Bias

Analogous to the broad class of tests for comparing survival time differences between the two arms of a study, results of the iLRT may yield biased results if censoring is related to prognosis or if survival probabilities are not stationary and instead depend upon when a participant is recruited into the clinical trial [88]. Likewise, the iLRT may experience a significant loss of power if competing risks are not independent or censoring is informative

(i.e., a correlation exists between censoring and the event of interest) [89]. Examples include drug withdrawal attributable to a lack of efficacy or intolerability. Furthermore, as a test of statistical significance, the iLRT is not designed to estimate the effect size for a treatment difference between groups or to compute confidence intervals of an effect [88].

While the objective of the iLRT is to reduce the false negative rate while achieving a statistically significant result, the procedure may experience a slight loss of power in the case of diminishing treatment effects, where the survival curves initially diverge but converge back together over time. If this is anticipated and the clinician has a specific interest in diminishing treatment effects, then the Max-Combo or FH (1,0) tests may represent a better choice for accommodating this possibility. When the curves extend beyond the point of diminishing treatment effect and then crossover, this poses interpretational challenges that may be best handled as a post hoc stratified analyses. The latter scenario merits exploring the underlying reasons for the crossing-over and any subgroup effects (e.g., potential treatment switching) before reaching any conclusions [7,90]. In the case of crossing hazards, a two-sample semiparametric procedure has been proposed as an alternative analytic approach [91]. Investigators also may consider the use a "standard of care reference arm" with a comparable hazard pattern.

A weighted LRT that is not consistent under stochastic ordering may not necessarily control the Type I error rate [92]. In Example 4, with the SRT as a comparison technique, we provide a heuristic argument that both the iLRT and default Max-Combo test independently control Type I error to within an absolute difference less than or equal to 0.0039 in the case of late separating survival curves, while preserving the false positive rate under proportional hazards (Example 3). Analogous to the consistent Prentice–Wilcoxon statistic, the weight for the iLRT is based upon the number of participants at risk for each time point. By taking the logarithm of the number at risk and scaling accordingly, the iLRT is bounded above by the Prentice–Wilcoxon test.

When censoring is not under the control of the investigator, censored participants may not have the same future risk of the outcome event as non-censored participants [93]. Consequently, there may not be a one-to-one correspondence between cause-specific hazard and cumulative incidence [94]. Such non-informative censoring can occur under competing risks and potentially bias risk estimates [95]. Unfortunately, commonly used methods to account for non-competing risks depend on the hazards being proportional, which may not always be the case when using the iLRT or other weighted procedures [96]. When appropriate, competing risks can be treated as random effects in a multilevel, mixed-effects model.

#### 5.4.2. Sparseness of Data and Small Sample Sizes

The iLRT may lack statistical power if few events accompany the divergence of treatment hazards or censoring is heavy [97]. Sparseness in the tails of the survival curves at the time of interim analysis also can hinder reliable sample size re-estimation. As asymptotic theory was used to establish limiting formulas, the small-sample behavior of the iLRT may be uncertain in such cases. With sparse data, bootstrapping or permutation methods may be considered for validating the model robustness of the iLRT.

#### 5.4.3. Computational Barriers

Standard available commercial software to compute power for weighted LRTs using the Lakatos–Cantor method generally are limited to a few weight options (e.g., standard log-rank, generalized Wilcoxon/Gehan–Breslow, and Tarone–Ware). However, a downloadable computer algorithm to compute the Lakatos–Cantor method for the iLRT and other user specified weights is available online [75].

### *5.5. Future Directions*

The basis of this manuscript relies on selected empirical examples to support the use of the iLRT. Other situations may necessitate a different approach, and future research

will help to delineate the most appropriate solution, such as adaptive or machine learning strategies [8,38,47]. The restricted mean survival times (RMSTs) method, which visually corresponds to the area under the Kaplan–Meier curve for a specified time period ($\tau$), is another method for analyzing non-constant hazards and may be useful as secondary analysis [45,98]. However, this technique depends on the arbitrary choice of ($\tau$). The misspecification of this value can yield statistically significant but clinically irrelevant results by focusing only on a particular region of the survival curves. Exploring an assortment of data-driven ($\tau$) points and accounting for these choices when estimating statistical significance is a promising area of ongoing research. Piecewise proportional hazard models also may be a good choice in some cases [99,100], and the hyperbolic cosine and logistic-like weight functions have received mention in the literature [52].

While a diverse array of weights and variance estimators have been proposed for the LRT, there is a paucity of comparative information regarding their versatility and efficiency under varying levels of non-proportionality, censoring, and competing risks [36,55,59,101]. Furthermore, when the event rate is low, weighted LRTs may not retain their range of flexibility [102]. Future analysis, beyond the scope of the current manuscript, may be merited.

## 6. Conclusions

A truly omnibus test is able to accurately detect survival differences over the clinical spectrum of a drug trial, regardless of whether a positive result is apparent from the start of therapy or only materializes later in the study (i.e., there is a time lag in the effectiveness of therapy). In contrast to the standard LRT, which treats all time points uniformly, an appropriately weighted LRT has the advantage of identifying significant delayed treatment effects with only a slight reduction in power for other survival outcomes. That is, under proportional hazards, with a nominal decrease in the probability of truly rejecting the null hypothesis, a substantial gain in efficiency for late separating survival curves is achieved [103].

While the quest for a "Holy Grail" test with infinite flexibility (i.e., immune to the type of non-proportional hazard) remains elusive, the single-weight iLRT possesses many of the desirable properties of such an omnibus method, particularly when the terminal event of one arm occurs before study completion. The iLRT equals or surpasses the default (four-component) Max-Combo method in many important applications and is objectively simple to implement with available computer code. The method does not require complex or timely simulations to estimate study power, and as a single-weight test, the iLRT does not involve implicit multiplicity correction nor depends on the arbitrary selection of weights. Nonetheless, in some cases, the iLRT may lack the flexibility and power of other more generalized multi-component Max-Combo tests or individual (two-parameter) Fleming–Harrington (FH) weights.

Relying entirely on a proportional hazards assumption when planning for and selecting a statistical test is unwise unless one is highly confident about the parallel shape of the ensuing hazard functions [98]. For example, the benefit of treatment may not occur immediately, but rather require a certain amount of time to overcome a lengthy disease period. A delayed treatment benefit also may be a consequence of "immunologic adjustment", which often occurs with certain newer-generation cancer drugs. In contrast, the antibiotic treatment of an infectious disease generally manifests a rapid treatment response.

The single-weight iLRT does not depend on an arbitrary choice of weights yet is relatively versatile and retains excellent power under delayed treatment effects. Nonetheless, a preponderance of investigators continue to use the more familiar assumption of constant event rates and proportional hazards in the design and analysis of randomized controlled trials, despite a potential loss of power and efficiency if this supposition does not hold [104].

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** No human subjects data were collected during the course of this study.

**Data Availability Statement:** Data for Example 1 is provided in Appendix B. Please refer to indicated references for data sources of the other examples.

**Acknowledgments:** The author would like to thank Robert A. Lew, Genevieve N. Dupuis, Sarah Milligan, Cynthia Hau, Kaitlin Cassidy, Maria Androsenko, and Charulata Jindal for their valuable comments and suggestions during the writing of the manuscript. Additionally, special thanks to Genevieve N. Dupuis for carefully checking the proofs and notation in the manuscript.

**Conflicts of Interest:** The author has no conflict of interests to declare. The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the United States Federal Government or the Department of Veterans Affairs.

## Appendix A. SAS Code for Inverse Log-Rank Test

```
proc sort data=a;
   by ti;
proc freq data=a;
   tables group censor/missprint;
proc lifetest data=a plots=s;
   time ti*censor(0);
    strata group;
proc sort data=a out=ti(keep=ti);
  by ti;
data ti;
   set ti;
      by ti;
         if first.ti;
proc freq data=a noprint;
   tables ti*censor/out=di1(keep=ti count rename=(count=di1));
      where group in (1) & censor in (1);
proc freq data=a noprint;
   tables ti*censor/out=di2(keep=ti count rename=(count=di2));
      where group in (2) & censor in (1);
data combo1(drop=j);
   merge ti(in=a) di1 di2;
      by ti;
         if a;
         array zero di1 di2;
            do j=1 to dim(zero);
               if zero{j}= . then zero{j}=0;
            end;
         di=sum(di1,di2);
ods listing close;
ods output "Product-Limit Estimates"=pl_all;
proc lifetest data=a;
   time ti*censor(0);
quit;
ods output close;
ods listing;
proc sort data=pl_all;
   by ti;
data ri (keep=left rename=(left=ri));
   set pl_all end=last;
      by ti;
         if last.ti & ^last;
data combo2;
   merge combo1(in=a) ri;
ods listing close;
ods output "Product-Limit Estimates"=pl_grp1;
proc lifetest data=a;
```

```
   time ti*censor(0);
      where group in (1);
quit;
ods output close;
ods listing;
proc sort data=pl_grp1;
   by ti;
data ri1 (keep=left rename=(left=ri1));
   set pl_grp1 end=last;
      by ti;
         if last.ti & ^last;
data ti1(keep=ti);
   set a(keep=group ti);
      by ti;
         where group in (1);
         if last.ti;
data ri1;
   merge ti1 ri1;
ods listing close;
ods output "Product-Limit Estimates"=pl_grp2;
proc lifetest data=a;
   time ti*censor(0);
      where group in (2);
quit;
ods output close;
ods listing;
proc sort data=pl_grp2;
   by ti;
data ri2 (keep=left rename=(left=ri2));
   set pl_grp2 end=last;
      by ti;
         if last.ti & ^last;
data ti2(keep=ti);
   set a(keep=group ti);
      by ti;
         where group in (2);
         if last.ti;
data ri2;
   merge ti2 ri2;
data combo3;
   retain ti di1 di2 di ri1 ri2 ri ei1 vi wi di1_sum di2_sum ei1_sum
          vi_sum top top_sum top_sum_sq chi_sq p;
   merge combo2(in=a) ri1 ri2;
      by ti;
         if a;
         if ri1= . & ri2> . & ri> . then ri1=ri-ri2;
```

```
            if ri2= . & ri1> . & ri> . then ri2=ri-ri1;
            di1_sum+di1;
            di2_sum+di2;
            ei1=di*ri1/ri;
            ei1_sum+ei1;
            wi=log(ri)/ri;
            if ri>1 then vi=(wi**2)*(ri1*ri2*di*(ri-di))/((ri**2)*(ri-1));
            if vi= . then vi=0;
            vi_sum+vi;
            top=wi*(di1-ei1);
            top_sum+top;
            top_sum_sq=top_sum**2;
            if vi_sum>0 then chi_sq=top_sum_sq/vi_sum;
            if chi_sq= . then chi_sq=0;
            p=1-probchi(chi_sq,1);
proc print data=combo3;
data combo4;
   set combo3(keep=p) end=last;
      if last;
proc print data=combo4;
run;
```

## Appendix B. Data Set for Example 1

```
data a;
   input group censor ti @@;
   cards;
   1 1 0.03559 1 1 0.03833 1 1 0.05202 2 1 0.06571 1 1 0.06845 1 1 0.08487 1 1 0.08487 2 1 0.09035
   2 1 0.09582 2 1 0.09582 1 1 0.09856 1 1 0.09856 1 1 0.10130 2 1 0.10404 1 1 0.10404 2 1 0.11225
   2 1 0.12320 1 1 0.12868 2 1 0.13142 1 1 0.13415 2 1 0.14511 1 1 0.15058 2 1 0.16427 1 1 0.16975
   1 1 0.16975 2 1 0.17248 1 1 0.17796 1 1 0.18617 1 1 0.19713 2 1 0.19986 1 1 0.21903 1 1 0.22450
   1 1 0.23546 1 1 0.24641 2 1 0.25462 2 1 0.26010 1 1 0.27379 2 1 0.28474 2 1 0.28747 2 1 0.29021
   2 1 0.29021 2 1 0.29569 1 1 0.30116 2 1 0.30116 2 1 0.30390 2 1 0.32033 2 1 0.32307 1 1 0.34223
   2 1 0.36687 1 1 0.36961 1 1 0.37782 2 1 0.38056 2 1 0.39151 1 1 0.39699 2 1 0.40520 1 1 0.40794
   2 1 0.41068 2 1 0.41068 2 1 0.43258 2 1 0.43532 2 1 0.44079 2 1 0.46270 2 1 0.47091 1 1 0.47091
   2 1 0.47639 2 1 0.48186 2 1 0.48460 2 1 0.49829 1 1 0.51472 2 1 0.51472 1 1 0.55578 2 1 0.56947
   2 1 0.60233 1 1 0.61328 1 1 0.61875 1 1 0.66256 1 1 0.67077 1 1 0.67351 1 1 0.68720 2 1 0.70637
   2 1 0.70637 1 1 0.70910 1 1 0.72005 2 1 0.73648 1 1 0.75291 1 1 0.75291 2 1 0.78850 2 1 0.83778
   1 1 0.84052 2 1 0.84052 1 1 0.85695 1 1 0.93634 2 1 0.93634 2 1 0.94182 2 1 1.02669 2 1 1.03491
   2 1 1.04312 2 1 1.05133 2 1 1.05955 2 1 1.08693 1 1 1.11157 1 1 1.14716 1 1 1.16632 1 1 1.21834
   2 1 1.22656 2 1 1.26215 2 1 1.26489 1 1 1.28131 1 0 1.35524 1 0 1.35797 2 1 1.39083 1 1 1.41547
   1 1 1.43737 1 1 1.54415 1 1 1.59069 2 1 1.61807 1 1 1.62628 2 1 1.70568 1 1 1.75770 1 1 1.76865
   2 1 1.83162 2 1 1.83984 2 1 2.26420 2 1 2.35729 2 1 2.42574 2 1 2.45038 2 1 2.46133 2 1 2.50513
   1 1 2.68857 2 1 2.78987 1 1 2.85010 1 1 2.88022 2 1 2.91581 2 1 2.98700 2 1 3.06913 1 1 3.28268
   1 0 3.31828 2 1 3.35661 1 0 3.61670 1 0 3.91786 2 1 4.03012 2 1 4.06845 2 1 4.09856 1 0 4.37235
   1 1 4.49008 2 1 4.50376 1 0 4.54483 1 0 4.75838 1 1 4.85695 1 0 4.87611 2 1 4.92266 2 1 5.13895
   1 1 5.24572 2 1 5.25394 2 1 5.88364 1 0 6.26694 1 0 6.34086
;
run;
```

## References


1. Ananthakrishnan, R.; Green, S.; Previtali, A.; Liu, R.; Li, D.; LaValley, M. Critical review of oncology clinical trial design under non-proportional hazards. *Crit. Rev. Oncol. Hematol.* **2021**, *162*, 103350. [CrossRef] [PubMed]
2. Fradet, Y.; Bellmunt, J.; Vaughn, D.J.; Lee, J.L.; Fong, L.; Vogelzang, N.J.; Climent, M.A.; Petrylak, D.P.; Choueiri, T.K.; Necchi, A.; et al. Randomized phase III KEYNOTE-045 trial of pembrolizumab versus paclitaxel, docetaxel, or vinflunine in recurrent advanced urothelial cancer: Results of >2 years of follow-up. *Ann. Oncol.* **2019**, *30*, 970–976. [CrossRef] [PubMed]
3. Ascierto, P.A.; Del Vecchio, M.; Robert, C.; Mackiewicz, A.; Chiarion-Sileni, V.; Arance, A.; Lebbé, C.; Bastholt, L.; Hamid, O.; Rutkowski, P.; et al. Ipilimumab 10 mg/kg versus ipilimumab 3 mg/kg in patients with unresectable or metastatic melanoma: A randomised, double-blind, multicentre, phase 3 trial. *Lancet Oncol.* **2017**, *18*, 611–622. [CrossRef] [PubMed]

4. Borghaei, H.; Paz-Ares, L.; Horn, L.; Spigel, D.R.; Steins, M.; Ready, N.E.; Chow, L.Q.; Vokes, E.E.; Felip, E.; Holgado, E.; et al. Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **2015**, *373*, 1627–1639. [CrossRef] [PubMed]
5. Schoenfeld, D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **1981**, *68*, 316–319. [CrossRef]
6. Wu, L.; Gilbert, P.B. Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics* **2002**, *58*, 997–1004. [CrossRef] [PubMed]
7. Public Workshop. *Oncology Clinical Trials in the Presence of Non-Proportional Hazards*; Duke-Margolis Health Policy Center: Washington, DC, USA, 2018.
8. Lee, S.-H. On the versatility of the combination of the weighted log-rank statistics. *Comput. Stat. Data Anal.* **2007**, *51*, 6557–6564. [CrossRef]
9. Fleming, T.R.; Harrington, D.P.; O'Sullivan, M. Supremum versions of the log-rank and generalized Wilcoxon statistics. *JASA* **1987**, *82*, 312–320. [CrossRef]
10. Peto, R.; Peto, J. Asymptotically efficient rank invariant test procedures. *J. R. Stat. Soc.* **1972**, *135*, 185–207. [CrossRef]
11. Cox, D.R.; Oakes, D. *Analysis of Survival Data*; CRC Press: Boca Raton, FL, USA, 1984.
12. Magirr, D. Non-proportional hazards in immuno-oncology: Is an old perspective needed? *Pharm. Stat.* **2021**, *20*, 512–527. [CrossRef]
13. Lindgren, B. *Statistical Theory*, 2nd ed.; The Macmillan Company: Toronto, ON, Canada, 1968.
14. Guenther, W. *Sampling Inspection in Statistical Quality Control*; Macmillian Publishing Co., Inc.: New York, NY, USA, 1977.
15. Collett, D. *Modelling Survival Data in Medical Research*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2015.
16. Fowobaje, K.; Wegbom, A.; Aboko, I.; Jolayemi, E.T. Testing the approximation of hypergeometric distribution by the binomial distribution. *IOSR J. Math.* **2016**, *12*, 10–16. [CrossRef]
17. Mood, A.; Graybill, F.; Boes, D. *Introduction to the Theory of Statistics*, 3rd ed.; McGraw-Hill Book Company: New York, NY, USA, 1974.
18. Woodroofe, M. *Probability with Applications*; Mc Graw-Hill, Inc.: New York, NY, USA, 1975.
19. Rivals, I.; Personnaz, L.; Taing, L.; Potier, M.C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **2007**, *23*, 401–407. [CrossRef] [PubMed]
20. Jäntschi, L. Formulas, algorithms and examples for binomial distributed data confidence interval calculation: Excess risk, relative risk and odds ratio. *Mathematics* **2021**, *9*, 2506. [CrossRef]
21. Bass, R.F.; Ruiz, P.A.; Baudoin, F.; Gordina, M.; Mariano, P.; Mostovyi, O.; Sengupta, A.; Teplyaev, A.; Valdez, E. *Upper Level Undergraduate Probability with Actuarial and Financial Applications*; University of Connecticut Department of Mathematics: Storrs, CT, USA, 2020.
22. Feller, W. *Introduction to Probability Theory and Its Application*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1968.
23. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **1966**, *50*, 163–170. [PubMed]
24. Prentice, R.L. Linear rank tests with right censored data. *Biometrika* **1978**, *65*, 167–179. [CrossRef]
25. O'Quigley, J. *Survival Analysis*; Springer: New York, NY, USA, 2021.
26. Lin, R.S.; León, L.F. Estimation of treatment effects in weighted log-rank tests. *Contemp. Clin. Trials Commun.* **2017**, *8*, 147–155. [CrossRef]
27. Mehrotra, K.G.; Michalek, J.E.; Mihalko, D. A relationship between two forms of linear rank procedures for censored data. *Biometrika* **1982**, *69*, 674–676. [CrossRef]
28. De-Mei, Y.; Li-Ran, W.; Lan, L. Conditional central limit theorems for a sequence of conditional independent random variables. *J. Korean Math. Soc.* **2014**, *51*, 1–15. [CrossRef]
29. Fleming, T.R.; Harrington, D.P. A class of hypothesis tests for one and two sample censored survival data. *Commun. Stat. Theory Methods* **1981**, *10*, 763–794. [CrossRef]
30. Garès, V.; Andrieu, S.; Dupuy, J.-F.; Savy, N. On the Fleming–Harrington test for late effects in prevention randomized controlled trials. *J. Stat. Theory Pract.* **2017**, *11*, 418–435. [CrossRef]
31. Wu, J. Sample size calculation for testing differences between cure rates with the optimal log-rank test. *J. Biopharm. Stat.* **2017**, *27*, 124–134. [CrossRef] [PubMed]
32. Ying, Z. Linear rank statistics for truncated data. *Biometrika* **1990**, *77*, 909–914. [CrossRef]
33. Rebolledo, R. Central limit theorems for local martingales. *Z. Für Wahrscheinlichkeitstheorie Und Verwandte Geb.* **1980**, *51*, 269–286. [CrossRef]
34. Dey, P.S.; Terlov, G. Stein's method for conditional central limit theorem. *Ann. Probab.* **2023**, *51*, 723–773. [CrossRef]
35. Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 16–21 June 1971; Volume 2, Probability Theory. pp. 583–602.
36. Lee, J.W. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* **1996**, *52*, 721–725. [CrossRef]

37. Lin, R.; Lin, J.; Roychoudhury, S.; Anderson, K.; Hu, T.; Huang, B.; Leon, L.; Liao, J.; Liu, R.; Luo, X.; et al. Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. *Stat. Biopharm. Res.* **2019**, *12*, 187–198. [CrossRef]
38. Yang, S.; Prentice, R. Improved logrank-type tests for survival data using adaptive weights. *Biometrics* **2010**, *66*, 30–38. [CrossRef] [PubMed]
39. Mukhopadhyay, P.; Ye, J.; Anderson, K.M.; Roychoudhury, S.; Rubin, E.H.; Halabi, S.; Chappell, R.J. Log-rank test vs max combo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in Immuno-oncology trials: A systematic review and meta-analysis. *JAMA Oncol.* **2022**, *8*, 1294–1300. [CrossRef]
40. Dormuth, I.; Liu, T.; Xu, J.; Yu, M.; Pauly, M.; Ditzhaus, M. Which test for crossing survival curves? A user's guideline. *BMC Med. Res. Methodol.* **2022**, *22*, 34. [CrossRef]
41. Wang, L.; Luo, X.; Zheng, C. A simulation-free group sequential design with max-combo tests in the presence of non-proportional hazards. *Pharm. Stat.* **2021**, *20*, 879–897. [CrossRef]
42. Prior, T.J. Group sequential monitoring based on the maximum of weighted log-rank statistics with the Fleming-Harrington class of weights in oncology clinical trials. *Stat. Methods Med. Res.* **2020**, *29*, 3525–3532. [CrossRef]
43. Magirr, D.; Jiménez, J.L. Design and analysis of group-sequential clinical trials based on a modestly weighted log-rank test in anticipation of a delayed separation of survival curves: A practical guidance. *Clin. Trials* **2022**, *19*, 201–210. [CrossRef]
44. Tsiatis, A.A. Repeated significance testing for a general class of statistics used in censored survival analysis. *JASA* **1982**, *77*, 855–861. [CrossRef]
45. Freidlin, B.; Korn, E.L. Methods for accommodating nonproportional hazards in clinical trials: Ready for the primary analysis? *J. Clin. Oncol.* **2019**, *37*, 3455–3459. [CrossRef] [PubMed]
46. Efird, J.T. *An Inverse Logarithmically-Scaled Rank Test Sensitive to Delayed Events*; Biometrics Section, The American Statistical Association: Alexandria, VA, USA, 1999; pp. 252–255.
47. Self, S.G. An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics* **1991**, *47*, 975–986. [CrossRef]
48. Cuzick, J. Asymptotic properties of censored linear rank tests. *Ann. Stat.* **1985**, *13*, 133–141. [CrossRef]
49. Efird, J.; Pardo, F. A non-parametric two-sample survival test based on a single occupancy fermi-dirac model for the discrete range distribution. In *Lifetime Data: Models in Reliability and Survival Analysis*; Jewell, N., Kimber, A., Lee, M., Whitmore, G., Eds.; Springer: Boston, MA, USA, 1996.
50. Knezevic, A.; Patil, S. Combination weighted log-rank tests for survival analysis with non-proportional hazards. In Proceedings of the SAS Global Forum, Washington, DC, USA, 29 March–1 April 2020.
51. Efird, J.T. Goldilocks rounding: Achieving balance between accuracy and parsimony in the reporting of relative effect estimates. *Cancer Inform.* **2021**, *20*, 1176935120985132. [CrossRef]
52. Qian, K.; Zhou, X. Weighted log-rank test for clinical trials with delayed treatment effect based on a novel hazard function family. *Mathematics* **2022**, *10*, 2573. [CrossRef]
53. Brown, B.W.; Hollander, M. *Statistics—A Biomedical Introduction*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
54. Neyman, J. Tests of statistical hypotheses which are unbiased in the limit. *Ann. Math. Stat.* **1938**, *9*, 69–86. [CrossRef]
55. Garès, V.; Andrieu, S.; Dupuy, J.-F.; Savy, N. A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests. *Electron. J. Statist.* **2014**, *8*, 841–860. [CrossRef]
56. Radhakrishna, S. Combination of results from several 2 × 2 contingency tables. *Biometrics* **1965**, *21*, 86–98. [CrossRef]
57. Tarone, R.E.; Ware, J. On distribution-free tests for equality of survival distributions. *Biometrika* **1977**, *64*, 156–160. [CrossRef]
58. Jennrich, R.I. Some exact tests for comparing survival curves in the presence of unequal right censoring. *Biometrika* **1984**, *71*, 57–64. [CrossRef]
59. Brown, M. On the choice of variance for the log rank test. *Biometrika* **1984**, *71*, 65–74. [CrossRef]
60. Finkelstein, D.M. A proportional hazards model for interval-censored failure time data. *Biometrics* **1986**, *42*, 845–854. [CrossRef] [PubMed]
61. Fay, M.P. Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics* **1996**, *52*, 811–822. [CrossRef]
62. Yang, S. Interim monitoring using the adaptively weighted log-rank test in clinical trials for survival outcomes. *Stat. Med.* **2019**, *38*, 601–612. [CrossRef]
63. Chiang, C.L. On regular best asymptotically normal estimates. *Ann. Math. Stat.* **1956**, *27*, 336–351. [CrossRef]
64. Efron, B.; Hinkley, D.V. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika* **1978**, *65*, 457–482. [CrossRef]
65. Chernoff, H.; Lehmann, E.L. The use of maximum likelihood estimates in $\chi 2$ tests for goodness of fit. *Ann. Math. Stat.* **1954**, *25*, 573–578. [CrossRef]
66. Ferguson, T. A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Ann. Math. Stat.* **1958**, *29*, 1046–1062. [CrossRef]
67. Neyman, J. Contribution to the theory of the $\chi 2$ test. In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1949; pp. 239–273.

68. Noether, G.E. On a theorem of Pitman. *Ann. Math. Statist.* **1955**, *26*, 64–68. [CrossRef]
69. Lehmann, E. Some comments on large sample tests. In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1949; pp. 451–457.
70. Zucker, D.M.; 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*, 853–864. [CrossRef]
71. Zucker, D.M. The efficiency of a weighed log-rank test under a percent error misspecification model for the log hazard ratio. *Biometrics* **1992**, *48*, 893–899. [CrossRef] [PubMed]
72. Ghahramani, M.; Sharafi, M.; Hashemi, R. Analysis of the progressively Type-II right censored data with dependent random removals. *J. Stat. Comp. Simul.* **2020**, *90*, 1001–1021. [CrossRef]
73. Efird, J.T.; Jindal, C. Using a counting process method to impute censored follow-up time data. *Int. J. Environ. Res. Public Health* **2018**, *15*, 690. [CrossRef]
74. Lakatos, E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* **1988**, *44*, 229–241. [CrossRef]
75. Cantor, A. *Survival Analysis Techniques for Medical Research*; SAS Publishing: Cary, NC, USA, 2003.
76. Lu, K. Sample size calculation for logrank test and prediction of number of events over time. *Pharm. Stat.* **2021**, *20*, 229–244. [CrossRef]
77. Aalen, O. Nonparametric inference for a family of counting processes. *Ann. Stat.* **1978**, *6*, 701–726. [CrossRef]
78. Cantor, A.B. Power calculation for the log rank test using historical data. *Control. Clin. Trials* **1996**, *17*, 111–116. [CrossRef]
79. Shih, J.H. Sample size calculation for complex clinical trials with survival endpoints. *Control. Clin. Trials* **1995**, *16*, 395–407. [CrossRef] [PubMed]
80. Harrell, F. *Regression Modeling Strategies with Application to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed.; Springer: New York, NY, USA, 2015.
81. Wainer, H.; Gessaroli, M.; Verdi, M. Visual revelations. Finding what is not there through the unfortunate binning of results: The mendel effect. *CHANCE* **2006**, *19*, 49–52. [CrossRef]
82. Freedman, D.; Diaconis, P. On the histogram as a density estimator:L2 theory. *Z. Für Wahrscheinlichkeitstheorie Und Verwandte Geb.* **1981**, *57*, 453–476. [CrossRef]
83. Sturges, H.A. The choice of a class interval. *JASA* **1926**, *21*, 65–66. [CrossRef]
84. Wand, M.P. Data-based choice of histogram bin width. *Am. Stat.* **1997**, *51*, 59–64. [CrossRef]
85. Scott, D.W. On optimal and data based histograms. *Biometrika* **1979**, *66*, 605–610. [CrossRef]
86. Doane, D.P. Aesthetic frequency classifications. *Am. Stat.* **1976**, *30*, 181–183. [CrossRef]
87. Bagust, A.; Beale, S.J. Exploring the effects of early censoring and analysis of clinical trial survival data on effectiveness and cost-effectiveness estimation through a case study in advanced breast cancer. *Med. Decis. Mak.* **2018**, *38*, 789–796. [CrossRef]
88. Bland, J.M.; Altman, D.G. The logrank test. *BMJ* **2004**, *328*, 1073. [CrossRef]
89. Williamson, P.R.; Kolamunnage-Dona, R.; Tudur Smith, C. The influence of competing-risks setting on the choice of hypothesis test for treatment effect. *Biostatistics* **2007**, *8*, 689–694. [CrossRef]
90. Jiménez, J.L.; Niewczas, J.; Bore, A.; Burman, C.F. A modified weighted log-rank test for confirmatory trials with a high proportion of treatment switching. *PLoS ONE* **2021**, *16*, e0259178. [CrossRef]
91. Yang, S.; Prentice, R. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika* **2005**, *92*, 1–17. [CrossRef]
92. Magirr, D.; Burman, C.F. Modestly weighted logrank tests. *Stat. Med.* **2019**, *38*, 3782–3790. [CrossRef] [PubMed]
93. Schuster, N.A.; Hoogendijk, E.O.; Kok, A.A.L.; Twisk, J.W.R.; Heymans, M.W. Ignoring competing events in the analysis of survival data may lead to biased results: A nonmathematical illustration of competing risk analysis. *J. Clin. Epidemiol.* **2020**, *122*, 42–48. [CrossRef] [PubMed]
94. Zhang, Z. Survival analysis in the presence of competing risks. *Ann. Transl. Med.* **2017**, *5*, 47. [CrossRef] [PubMed]
95. Austin, P.C.; Lee, D.S.; Fine, J.P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **2016**, *133*, 601–609. [CrossRef] [PubMed]
96. Fine, J.P.; Gray, R.J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **1999**, *94*, 496–509. [CrossRef]
97. Pepe, M.S.; Fleming, T.R. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics* **1989**, *45*, 497–507. [CrossRef]
98. Freidlin, B.; Korn, E.L. Reply to H. Uno et al. and B. Huang et al. *J. Clin. Oncol.* **2020**, *38*, 2003–2004. [CrossRef]
99. Yu, C.; Huang, X.; Nian, H.; He, P. A weighted log-rank test and associated effect estimator for cancer trials with delayed treatment effect. *Pharm. Stat.* **2021**, *20*, 528–550. [CrossRef]
100. Liu, S.; Chu, C.; Rong, A. Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. *Pharm. Stat.* **2018**, *17*, 541–554. [CrossRef]
101. Garès, V.; Andrieu, S.; Dupuy, J.F.; Savy, N. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Stat. Med.* **2015**, *34*, 541–557. [CrossRef] [PubMed]
102. Buyske, S.; Fagerstrom, R.; Ying, Z. A class of weighted log-rank tests for survival sata when the event is rare. *JASA* **2000**, *95*, 249–258. [CrossRef]

103. Su, Z.; Zhu, M. Is it time for the weighted log-rank test to play a more important role in confirmatory trials? *Contemp. Clin. Trials Commun.* **2018**, *10*, A1–A2. [CrossRef]
104. Jachno, K.; Heritier, S.; Wolfe, R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. *BMC Med. Res. Methodol.* **2019**, *19*, 103. [CrossRef] [PubMed]