

Estimation of the relative risk following group sequential procedure based upon the weighted log-rank statistic*

Grant Izmirlan

*National Cancer Institute; Executive Plaza North, Suite 3131
6130 Executive Blvd, MSC 7354; Bethesda, MD 20892-7354
e-mail: izmirlig@mail.nih.gov*

Abstract: In this paper we consider a group sequentially monitored trial on a survival endpoint, monitored using a weighted log-rank (WLR) statistic with deterministic weight function. We introduce a summary statistic in the form of a weighted average logged relative risk and show that if there is no sign change in the instantaneous logged relative risk, there always exists a bijection between the WLR statistic and the weighted average logged relative risk. We show that this bijection can be consistently estimated at each analysis under a suitable shape assumption, for which we have listed two possibilities. We indicate how to derive a design-adjusted p-value and confidence interval and suggest how to apply the bias-correction method. Finally, we document several decisions made in the design of the NLST interim analysis plan and in reporting its results on the primary endpoint.

AMS 2000 subject classifications: Primary 62L12, 62L12; secondary 62N02.

Keywords and phrases: Weighted Logrank Statistic, Group Sequential, Interim Analysis, Estimation.

1. Introduction

Time to event, e.g. disease specific mortality, is the primary endpoint in many clinical trials. The use of group sequential boundaries in monitoring the trial is not only commonplace, but ethically mandated in all trials of human subjects. The logrank statistic is often the monitoring statistic of choice due to its natural connection with the relative risk, which is often the parameter of inference. This natural connection, which is based upon the assumption of proportional hazards, admits a one-to-one correspondence between the inferential procedure based upon the usual standard normal scale and that based on the scale of the natural parameter. However, the assumption of proportional hazards is not always a reasonable assumption. In many subject areas, e.g. in disease-prevention trials, one expects that the hazard ratio will not be constant. Much of the prior work on the use of the weighted logrank statistic in a sequential design is confined to the use a weighting function from the $G^{\rho,\gamma}(t) = S^\rho(t)(1 - S(t))^\gamma$ family, of Fleming and Harrington, [2]. They suggest two major types of problems which can arise. First, they argue that use of the weighted logrank statistic does not

*This article is a U.S. Government work and is in the public domain in the U.S.A.

reproduce the single point analysis in the way that is desired. Most notably, they argue, there is no clinically meaningful parameter that allows the values of the monitoring statistic and sequential boundaries to be cast into a clinically meaningful scale. They believe that this problem is further aggravated when the range of the weighting function over the duration of the trial is quite large, such as is the case with the $G^{0,1}$ weight function (Gillen and Emerson, [4]) and suggest a re-weighting scheme whereby the most weight is given to the most recent data collected at each analysis. Secondly, they argue that if the chosen weighting function is non-deterministic or trial-specific then it is impossible to compare results from different clinical trials, (Gillen and Emerson, [3, 5]). While the bulk of these cautious remarks are useful to know in their own right, several important points have been omitted from the discussion. Firstly, as we will show, there is a natural, clinically meaningful parameter, the weighted average logged relative risk, that is connected bijectively to the weighted logrank statistic when there is no change in sign in the instantaneous logged relative risk. Under suitable shape assumptions, the bijection can be estimated at each analysis. We will show that the asymptotic distribution of the WLR statistic, suitably normalized is a Brownian motion plus drift under nothing but boundedness conditions. In two corollaries, we demonstrate how each of two presented shape assumptions translates into a form of the drift function and consequently, into an estimator of the weighted average logged relative risk. We then demonstrate how the usual results concerning monitoring and end of trial estimation follow. Finally, we note that this bijection between the weighted logrank statistic and the weighted average logged relative risk allows the values of the monitoring statistic, efficacy and futility boundaries, and reported point estimate and confidence interval to be cast into a clinically meaningful scale.

2. Terminology and framework

We consider a two armed randomized trial of the effect of an intervention upon a time to event that is run until time τ . Let \tilde{T}_i be the possibly unobserved time to event and let C_i a right censoring time. We assume non-informative censoring for simplicity. Let $T_i = \tilde{T}_i \wedge C_i$ be the observed time on study and let $\delta_i = I(\tilde{T}_i \leq C_i)$ be the event indicator. Let X_i indicates membership in the intervention arm ($X_i = 1$) or control arm ($X_i = 0$). We assume, conditional upon X_i , that individuals, $i = 1, \dots, n$ are distributed independently and identically. Let $dH_0(t)$ and $dH_1(t)$ be the trial arm specific cumulative hazard increments. We assume throughout that $H_0(t)$ is finite for all t on $[0, \tau]$. For the instantaneous logged hazard ratio, we write

$$\beta(t) = \log \left\{ \frac{dH_1(t)}{dH_0(t)} \right\}. \quad (2.1)$$

Let $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $dN_i(t) = N_i(t) - N_i(t-)$ be the subject level counting process and its increments, respectively. Let $N_n(t) = \sum_i N_i(t)$

and $dN_n(t) = N_n(t) - N_n(t-)$ be the aggregated counting process and its increments, respectively. Note that the following difference is a compensated counting process martingale:

$$dM_i(t) = dN_i(t) - I(T_i \geq t) \exp(X_i \beta(t)) dH_0(t) \quad (2.2)$$

Let $E_n(t, 0) = \sum_i X_i I(T_i \geq t) / \sum_i I(T_i \geq t)$ denote the proportion of the population at risk at time t in the intervention arm, and let $e(t, 0) = \lim_{a.s. n \rightarrow \infty} E_n(t, 0)$ and let $G(t) = \lim_{a.s.} dN_n(t)/n$. Let $\mathcal{IF}_n(t) = \int_0^t E_n(\xi, 0)(1 - E_n(\xi, 0)) dN_n(\xi)/n$ and let $\mathcal{IF}(t) = \int_0^t e(\xi, 0)(1 - e(\xi, 0)) dG(\xi)$. We introduce the following notation for cross moment integrals against $d\mathcal{IF}$ over $(0, t)$:

$$\langle \psi_1 | \mathcal{IF} | \psi_2 \rangle_t = \int_0^t \psi_1(\xi) \psi_2(\xi) d\mathcal{IF}(\xi). \quad (2.3)$$

For reasons that will become clear below, we consider the target of our investigation to be the following weighted average logged relative risk:

$$\beta^* = \frac{\langle Q | \mathcal{IF} | \beta \rangle_\tau}{\langle Q | \mathcal{IF} | 1 \rangle_\tau}. \quad (2.4)$$

Let $q(t) = \beta(t)/\beta^*$. This provides a representation of the instantaneous logged relative risk function, $\beta(t) = \beta^* q(t)$ as the product of its weighted average value, β^* times a shape function, q . Note it follows that the shape function has weighted average value equal to 1:

$$1 = \frac{\langle Q | \mathcal{IF} | q \rangle_\tau}{\langle Q | \mathcal{IF} | 1 \rangle_\tau}. \quad (2.5)$$

At follow-up time t , the \sqrt{n} normalized score statistic with weighting function Q is:

$$U_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t Q(\xi) \{X_i - E_n(\xi, 0)\} dN_i(\xi). \quad (2.6)$$

Its estimated variance is:

$$V_n(t) = \frac{1}{n} \int_0^t Q^2(\xi) E_n(\xi, 0) (1 - E_n(\xi, 0)) dN_n(\xi) = \langle Q | \mathcal{IF}_n | Q \rangle_t. \quad (2.7)$$

Let $v(t) = \lim_{a.s.} V_n(t)$. Note that $v(t) = \langle Q | \mathcal{IF} | Q \rangle_t$. Let $f_n(t; \tau) = V_n(t)/V_n(\tau)$ and $f(t; \tau) = v(t)/v(\tau)$. We will on occasion use the shorthand $f_{n,j}$ and f_j for $f_n(t; \tau)$ and $f(t; \tau)$, respectively. Also, let $m_n(t) = \langle Q | \mathcal{IF}_n | Q \rangle_t$ and $m(t) = \langle Q | \mathcal{IF} | Q \rangle_t$. We consider the weighted log-rank (WLR) statistic at time t on several ‘‘scales’’

- (i) The standard normal scale: $Z_n(t) = U_n(t)/\sqrt{V_n(t)}$
- (ii) The ‘‘Brownian scale’’: $X_n(t) = U_n(t)/\sqrt{V_n(\tau)}$

3. Main Result

Condition 3.1. *The instantaneous logged relative risk function, β , is bounded on $[0, \tau]$.*

Condition 3.2. *The chosen weighting function, Q , is bounded on $[0, \tau]$ and deterministic.*

Recall that a weighting functions is always non-negative. The stipulated boundedness in conditions 3.1 and 3.2 above can be relaxed to being of class L^2 with respect to the measure $d\mathbb{F}$, as this is all that is really required. While the context will involve monitoring the statistic at a sequence of interim analyses, for the time being, we suppress this aspect and consider instead the following more general and generic result which holds under the weakest set of assumptions:

Theorem 3.1. *Under conditions 3.1 and 3.2, then under the family of local alternatives, $\beta_n^* = b^*/\sqrt{n}$, the score statistic, normalized to the “Brownian scale” is asymptotically a Brownian motion on $[0, 1]$ plus a drift.*

$$X_n(t) \xrightarrow{\mathcal{D}} W(f(t; \tau)) + \mu(t) \quad (3.1)$$

where the “time scale” for the Brownian motion is the variance ratio or information fraction, $f(t; \tau) = v(t)/v(\tau)$, and the drift, parameterized by t is

$$\mu(t) = \frac{\langle Q|\mathbb{F}|q \rangle_t}{\sqrt{\langle Q|\mathbb{F}|Q \rangle_\tau}} b^*. \quad (3.2)$$

The proof of 3.1 is given in appendix 8.1. Notice, first, that from equations 2.5 and 3.2, it follows that the value of the drift function at the scheduled end of the trial is

$$\mu(\tau) = \frac{\langle Q|\mathbb{F}|1 \rangle_\tau}{\sqrt{\langle Q|\mathbb{F}|Q \rangle_\tau}} b^*. \quad (3.3)$$

Thus, without any additional assumptions on the shape function, q , we have the following corollary:

Corollary 3.1. *At the planned conclusion of the trial, τ , an estimate of β^* is given by the following:*

$$\widehat{\beta}^* = X_n(\tau) \frac{\sqrt{\langle Q|\mathbb{F}_n|Q \rangle_\tau}}{\sqrt{n} \langle Q|\mathbb{F}_n|1 \rangle_\tau}. \quad (3.4)$$

- (i) $\widehat{\beta}^*$ is unbiased
- (ii) An estimate of its variance is given by

$$\text{var} \left[\widehat{\beta}^* \right] = \frac{\langle Q|\mathbb{F}_n|Q \rangle_\tau}{n \langle Q|\mathbb{F}_n|1 \rangle_\tau^2}. \quad (3.5)$$

4. Estimates of β^* in a Trial Stopped Early

Obtaining an estimate of β^* at a trial stopped early due to an efficacy boundary crossing will require more assumptions on the shape function, q . At a minimum in order to have a monotone drift function which is necessary for proper monitoring, we require the following.

Condition 4.1. *The shape function, q , is non-negative.*

Since the drift's function's dependence on t is through an integral of a non-negative function, we have the following corollary:

Corollary 4.1. *If conditions 3.1, 3.2 and 4.1 are true then the conclusion of theorem 3.1 holds and the drift function is monotone increasing or decreasing in t , depending upon the sign of b^* .*

Note also that as the inverse of an increasing function is also increasing, the drift function can also be considered a monotone function of the information fraction. This would, of course, lead to a natural estimate of β^* in a trial stopped early except for the fact that we have no knowledge of q . In order to have a more useful estimator for β^* in trials stopped early, we opt for a semi-parametric model. In the following, we list two possibilities. The most natural shape condition to impose is true if our choice of weight function was the optimal one among all possible choices.

Condition 4.2. *The shape function, q , is proportional to our chosen weighting function, $q(t) = K Q(t)$.*

Note that as the weighted average of the shape function must equal 1 as in equation 2.5 it follows that the constant of proportionality, K , must be

$$K = \frac{\langle Q|\mathbb{F}|1 \rangle_\tau}{\langle Q|\mathbb{F}|Q \rangle_\tau}. \quad (4.1)$$

Corollary 4.2. *If conditions 3.1, 3.2 and 4.2 are true then*

- (i) X_n is asymptotically a Brownian motion with a drift that is linear in the information fraction:

$$\mu(t) = \frac{\langle Q|\mathbb{F}|1 \rangle_\tau}{\sqrt{\langle Q|\mathbb{F}|Q \rangle_\tau}} f(t; \tau) b^*. \quad (4.2)$$

- (ii) *If the trial is stopped at an analysis number J at calendar time t_J due to an efficacy boundary crossing, then we have the following estimate of β^**

$$\hat{\beta}^* = \frac{X_n(t_J)}{f_n(t_J; \tau)} \frac{\sqrt{\langle Q|\mathbb{F}_n|Q \rangle_\tau}}{\sqrt{n} \langle Q|\mathbb{F}_n|1 \rangle_\tau} \quad (4.3)$$

- (iii) *An estimate of the mean-squared error is given by:*

$$\text{mse} \left[\hat{\beta}^* \right] = \frac{\langle Q|\mathbb{F}_n|Q \rangle_\tau}{n f_n(t_J; \tau) \langle Q|\mathbb{F}_n|1 \rangle_\tau^2} \quad (4.4)$$

Another natural shape condition is true when we have opted for a weighted statistic but the true shape is constant.

Condition 4.3. *The shape function, q , is identically 1.*

Corollary 4.3. *If conditions 3.1, 3.2 and 4.3 are true then*

(i) X_n is asymptotically a Brownian motion the following drift:

$$\mu(t) = \frac{\langle Q|\mathbb{F}|1\rangle_\tau}{\sqrt{\langle Q|\mathbb{F}|Q\rangle_\tau}} r(t, \tau) b^*, \quad (4.5)$$

where $r(t; \tau) = \langle Q|\mathbb{F}|1\rangle_t / \langle Q|\mathbb{F}|1\rangle_\tau$, which is an increasing function of t and takes the values 0 at $t = 0$ and 1 at $t = \tau$.

(ii) If the trial is stopped at an analysis number J at calendar time t_J due to an efficacy boundary crossing, then we have the following estimate of β^*

$$\hat{\beta}^* = \frac{X_n(t_J)}{r_n(t_J; \tau)} \frac{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}{\sqrt{n} \langle Q|\mathbb{F}_n|1\rangle_\tau}, \quad (4.6)$$

where $r_n(t; \tau) = \langle Q|\mathbb{F}_n|1\rangle_t / \langle Q|\mathbb{F}_n|1\rangle_\tau$

(iii) An estimate of the mean-squared error is given by:

$$\text{mse} \left[\hat{\beta}^* \right] = \frac{f_n(t_J; \tau) \langle Q|\mathbb{F}_n|Q\rangle_\tau}{n r_n(t_J; \tau)^2 \langle Q|\mathbb{F}_n|1\rangle_\tau^2} \quad (4.7)$$

5. Application to Monitoring and Final Reporting in a Clinical Trial

The relationship between the drift of the WLR statistic and the weighted average logged relative risk parameter provided by theorem 3.1 and its corollaries can be used in the monitoring and final reporting of a clinical trial.

5.1. Futility Boundary

Our comments regarding monitoring a trial are made within the context of boundaries constructed using the Lan-Demets procedure, [6]. Construction of the efficacy boundary is done under the null hypothesis that the drift function is identically zero and can be done without appealing to the results presented here. If a futility boundary is specified in the design then under either of the shape assumptions, one can apply the corresponding corollary 4.2 or corollary 4.3 to calculate the drift function at each interim analysis which is required to compute the futility boundary under the Lan-Demets approach [6]. Note that the shape assumption being made must be part of the interim analysis plan design. In the following discussion we will assume that the optimal weighting shape condition 4.2 was specified in the design so that the discussion focuses on the application of corollary 4.2. In this case, β^* is the weighted average logged relative risk for which the study is powered to detect and must also be specified in the interim

analysis plan design. The values of $v(\tau) = \langle Q|IF|Q \rangle_\tau$ and $m(\tau) = \langle Q|IF|1 \rangle_\tau$ at the planned termination of the study, τ , must also be specified in the interim analysis plan design. We demonstrate in appendix 8.2 when the only source of censoring is administrative censoring or other cause mortality, how these functionals can be projected for a specific choice of weighting function, Q , based upon projected values of the cross-arm pooled cumulative hazard function at several landmark times on study. We remark here that following consensus, we recommend using a non-binding futility boundary which is constructed after construction of an efficacy boundary which ignores the existence of the futility boundary. This is preferred to the joint construction of efficacy and futility boundaries as that approach results in a discounted efficacy criterion.

5.2. Prediction at End of Trial

When the trial is stopped at an efficacy or futility boundary crossing, or at the scheduled end of the trial, and if the optimal weighting shape assumption 4.2 was specified in the design, then corollary 4.2 can be used to convert the value of the WLR statistic on the Brownian scale, $X_n(t_j)$, to an estimate of the weighted average logged relative risk, $\hat{\beta}^*$. Therefore, our point estimate is

$$\hat{\beta}^* = \frac{X_n(t_j)}{f_{n,j}} \frac{\sqrt{\langle Q|IF_n|Q \rangle_\tau}}{\sqrt{n} \langle Q|IF_n|1 \rangle_\tau} \quad (5.1)$$

We use the values of $v(\tau) = \langle Q|IF|Q \rangle_\tau$ and $m(\tau) = \langle Q|IF|1 \rangle_\tau$ which are specified in the interim analysis plan design. As mentioned above, when it is obtained at an efficacy boundary crossing, these type of estimates are known to be biased away from the null (see e.g. Liu and Hall, [7]). The construction of a design-adjusted confidence interval and adjustment of this estimate for the above mentioned bias are standard results, especially under the optimal weighting shape condition 4.2 which leads, in corollary 4.2, to a drift that is linear in the information fraction. For sake of completeness, we outline below how to compute a design adjusted p-value, construct a design-adjusted confidence interval and how to calculate the bias adjusted estimate of the weighted average logged relative risk. All three of these tasks involve the sampling density under the null hypothesis of the sufficient statistic, $(J, X_n(t_J))$, where J and $X_n(t_J)$ are the analysis number and the value of the weighted logrank statistic at an efficacy crossing. The sampling density of $(J, X_n(t_J))$ takes the following form. First, for $j = 1$, $\pi((1, x)) = \mathbb{P}\{X_n(t_1) = x\}$. For $j > 1$,

$$\begin{aligned} \pi((j, x) ; \mathbf{b}_{1:(j-1)}, \mathbf{f}_{1:j}) & \quad (5.2) \\ &= \frac{d}{dx} \mathbb{P}_{H_0} \{J = j \text{ and } X_n(t_\ell) < \sqrt{f_\ell} b_\ell, \ell = 1, \dots, j-1, X_n(t_j) = x\} \end{aligned}$$

Here $\mathbf{b}_{1:(j-1)}$ is the sequence of efficacy boundary points at all prior analyses and $\mathbf{f}_{1:j}$ is the sequence of information fractions at all analyses prior and current. In

the following $\mathbf{b}_{1:\ell}$ and $\mathbf{f}_{1:\ell}$ for $\ell < 1$ denote the empty sequence. The construction and form of this density is reviewed in appendix 8.3. Let

$$\bar{\Pi}((j, x); \mathbf{b}_{1:(j-1)}, \mathbf{f}_{1:j}) = \int_x^\infty \pi((j, \xi); \mathbf{b}_{1:(j-1)}, \mathbf{f}_{1:j}) d\xi \quad (5.3)$$

be the joint probability under π that $J = j$ and $X_n(t_j)$ is in the right tail (x, ∞) . In order to calculate a p-value and construct a confidence interval which account for the sequential design, we must choose an ordering of the sample space for the statistic $(J, X_n(t_J))$. Here we prefer to use the following ordering: $(j, x) > (k, y)$ if and only if $(j = k \text{ and } x > y)$ or $j < k$. This ordering is applicable when the rejection region is convex, as is the case with Lan-Demets boundaries constructed using a smooth spending function. The discussion of the p-value and of the confidence interval is in the setting of symmetric 2-sided boundaries and when sign of the alternative hypothesis is positive as it is a simple matter to apply these results to the case where the sign of the alternative hypothesis is negative.

P-value

Under the ordering given above, the region further away from the null than $(J, X_n(t_J))$ is the union of all prior rejection regions with the right tail at $X_n(t_J)$. Thus the design-adjusted or sequential p-value is:

$$\bar{\Pi}((J, X_n(t_J)); \mathbf{b}_{1:(J-1)}, \mathbf{f}_{1:J}) + \sum_{\ell=1}^{J-1} \bar{\Pi}((\ell, b_\ell); \mathbf{b}_{1:\ell-1}, \mathbf{f}_{1:\ell}), \quad (5.4)$$

Confidence Interval

If the probability of type one error that remained prior to analysis J is $\alpha_{tot} - \alpha_{J-1}$ then a two sided design-adjusted confidence interval for $\hat{\beta}^*$ is derived as follows. If we denote by x_u the solution in x of the equation

$$\alpha_{tot} - \alpha_{J-1} = \bar{\Pi}((J, x); \mathbf{b}_{1:(J-1)}, \mathbf{f}_{1:J}) + \sum_{\ell=1}^{J-1} \bar{\Pi}((\ell, b_\ell); \mathbf{b}_{1:\ell-1}, \mathbf{f}_{1:\ell}), \quad (5.5)$$

then the design-adjusted confidence interval is

$$\hat{\beta}^* \pm \frac{x_u}{\sqrt{f_{n,J}}} \sqrt{\text{mse} [\hat{\beta}^*]}, \quad (5.6)$$

where $\text{mse} [\hat{\beta}^*]$ is the estimated mean-squared error of $\hat{\beta}^*$ as given in part (iii) of corollary 4.2. Note that when the efficacy boundary is one-sided one can still

construct a 2-sided confidence interval by replacing $\alpha_{tot} - \alpha_{J-1}$ above with $1/2$ its value.

Bias Adjustment

As in Liu and Hall, [7], bias adjustment is done recursively as follows. First,

$$\tilde{\zeta}(1, x) = \frac{x}{f_1} \quad (5.7)$$

Continuing,

$$\tilde{\zeta}(j, x) = \int_{-\infty}^{\sqrt{f_j} b_j} \tilde{\zeta}(j-1, \xi) \pi((j-1, \xi); \mathbf{b}_{1:(j-1)}, \mathbf{f}_{1:(j-1)}) \phi_{\Delta_j}(x - \xi) d\xi \quad (5.8)$$

The bias adjusted estimate, $\tilde{\beta}^*$, of the weighted average logged relative risk, β^* , is obtained by replacing $X_n(t_J)/f_{n,J}$ in part (ii) of corollary 4.2 with $\tilde{\zeta}(J, X_n(t_J))$ to obtain the following:

$$\tilde{\beta}^* = \tilde{\zeta}(J, X_n(t_J)) \frac{\sqrt{\langle Q | \mathcal{F}_n | Q \rangle_\tau}}{\sqrt{n} \langle Q | \mathcal{F}_n | 1 \rangle_\tau} \quad (5.9)$$

The design-adjusted confidence interval is the same as given above, but now centered about $\tilde{\beta}^*$

$$\tilde{\beta}^* \pm \frac{x_u}{\sqrt{f_{n,J}}} \sqrt{\text{mse}[\hat{\beta}^*]}, \quad (5.10)$$

6. The NLST

The design of the National Lung Screening Trial (NLST) [8] interim analysis plan stipulated a one-sided efficacy boundary constructed using the Lan-Demets procedure with a total probability of type one error set to 0.05. The trial had 90% power to detect a relative risk of 0.79 at a sample size of 25,000 per arm, accounting for contamination and non-compliance that could attenuate this effect to 0.85. The trial began randomization on August 5th, 2002 and concluded randomization on April 26th, 2004. A non-binding futility boundary was used. The drift was derived under the optimal weighting shape assumption, 4.2, and incorporated the design alternative $\beta^* = \log(0.85)$. Initial estimates of $v(\tau)$ and $m(\tau)$ were posed in the design. These were updated by using a least squares quadratic curve to project required future values of H as data accumulated. During the run of the trial, projected values of the end of trial functionals $v(\tau)$ and $m(\tau)$ did not vary more than $\pm 5\%$. Interim analyses occurred starting in Spring of 2006 and continued annually until the 5th analysis. The 6th analysis occurred 6 months after the 5th. Data on the primary endpoint was backdated roughly 18 months to allow more complete ascertainment by the endpoint verification team. The efficacy boundary was crossed at the sixth interim analysis,

using data backdated to January 15th 2009. Data on the primary endpoint was collected only for events occurring through December 31, 2009 so this was used as the scheduled termination date. The raw estimated weighted logged relative risk and its design-adjusted confidence interval were derived. The bias adjusted weighted logged relative risk was compared to the raw estimate. As the raw estimate is asymptotically unbiased, and since the crude risk ratio is the most straightforward and tangible summary of the trial results, the trial leadership decided to report the crude risk ratio together with the exponentiated raw estimate's design-adjusted confidence interval.

7. Discussion

We have shown that there is a natural clinically meaningful parameter, the weighted average logged relative risk, that is connected to the weighted logrank statistic. When $\beta(t)$ does not change sign, the connection is a bijection. We have shown that under suitable shape assumptions, this bijection can be estimated at each analysis. We have shown how this bijection between the weighted logrank statistic and the weighted average logged relative risk allows the values of the monitoring statistic, efficacy and futility boundaries, and reported point estimate and confidence interval to be cast into a clinically meaningful scale. We have indicated how to derive a design-adjusted p-value and confidence interval and how bias adjustment of the estimate may be done using known methods. Finally, we have documented several decisions made in the design of the NLST interim analysis plan and in reporting its results on the primary endpoint.

References

- [1] ARMITAGE, P., MCPHERSON, C. K. and ROWE, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* **132** 235–244.
- [2] FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.
- [3] GILLEN, D. and EMERSON, S. (2005). Information growth in a family of weighted logrank statistics under interim analyses. *Sequential Analysis* **24** 1–22.
- [4] GILLEN, D. and EMERSON, S. (2005). A note on P-values under group sequential testing and nonproportional hazards. *Biometrics* **61** 546–551.
- [5] GILLEN, D. and EMERSON, S. (2007). Non-transitivity in a class of weighted logrank statistics under nonproportional hazards. *Statistics and Probability Letters* **77** 123–130.
- [6] LAN, K. K. G. and DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663.
- [7] LIU, A. and HALL, W. J. (1999). Unbiased estimation following a group sequential test. *Biometrika* **86** 71–78.

- [8] THE NATIONAL LUNG SCREENING TRIAL RESEARCH TEAM, (2011). The National Lung Screening Trial: Overview and Study Design. *Radiology* **258** 243–253.

8. Appendices

8.1. Proof of Theorem 3.1

We follow the usual method of adding and subtracting the differential of the compensator, and thereby express U_n as a sum of a term that is asymptotically mean zero Gaussian process and a drift function which grows as \sqrt{n} .

$$\begin{aligned}
U_n(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t Q(\xi) \{X_i - E_n(\xi, 0)\} dM_i(\xi) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t Q(\xi) \{X_i - E_n(\xi, 0)\} I(T_i \geq \xi) \exp(X_i q(\xi) \beta^*) dH_0(\xi) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t Q(\xi) \{X_i - E_n(\xi, 0)\} dM_i(\xi) \\
&\quad + \sqrt{n} \int_0^t Q(\xi) \{E_n(\xi, \beta^*) - E_n(\xi, 0)\} R_n(\xi, \beta^*) dH_0(\xi), \tag{8.1}
\end{aligned}$$

where in the above, $R_n(\xi, \beta^*) = 1/n \sum_i I(T_i \geq \xi) \exp(X_i q(\xi) \beta^*)$, and $E_n(\xi, \beta^*) = 1/(n R_n(\xi, \beta^*)) \sum_i X_i I(T_i \geq \xi) \exp(X_i q(\xi) \beta^*)$.

By linearizing the difference, $E_n(\xi, \beta^*) - E_n(\xi, 0)$ about $\beta^* = 0$ we obtain

$$\begin{aligned}
U_n(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t Q(\xi) \{X_i - E_n(\xi, 0)\} dM_i(\xi) \\
&\quad + \sqrt{n} \beta^* \int_0^t Q(\xi) q(\xi) E_n(\xi, 0) \{1 - E_n(\xi, 0)\} R_n(\xi, \beta^*) dH_0(\xi). \tag{8.2}
\end{aligned}$$

We normalize by $\sqrt{V_n(\tau)}$ and replace the differential $R_n(\xi, \beta^*) dH_0(\xi)$ with $dN_n(\xi)/n$. The latter is possible because integrals of bounded functions against the difference of the differentials are consistent to zero.

$$\begin{aligned}
X_n(t) &= \frac{1}{\sqrt{n V_n(\tau)}} \sum_{i=1}^n \int_0^t Q(\xi) \{X_i - E_n(\xi, 0)\} dM_i(\xi) \\
&\quad + \sqrt{\frac{n}{V_n(\tau)}} \beta^* \int_0^t Q(\xi) q(\xi) E_n(\xi, 0) \{1 - E_n(\xi, 0)\} \frac{dN_n(\xi)}{n} \\
&= W_n(f_n(t; \tau)) + \frac{\langle Q | \mathbb{F}_n | q \rangle_t}{\sqrt{\langle Q | \mathbb{F}_n | Q \rangle_\tau}} \sqrt{n} \beta^*. \tag{8.3}
\end{aligned}$$

The first term is easily recognized to be asymptotic in distribution to a standard Brownian motion. The reader can either directly apply Robolledo's martingale central limit theorem, verifying that in the case that integrands and intensities are bounded all conditions are satisfied, or apply a more direct result, such as theorem (6.2.1) in Fleming and Harrington [2]. Under the family of local alternatives, $\beta_n^* = b^*/\sqrt{n}$, then by the comments following expression 8.2, the second term is easily seen to be consistent to the drift function listed in expression 3.2. Therefore the result follows by Slutsky's theorem.

8.2. End of Trial Functionals

In this section we demonstrate how to project values of the variance $v(\tau) = \langle Q|F|Q \rangle_\tau$, and the "first moment" $m(\tau) = \langle Q|F|1 \rangle_\tau$ at the scheduled end of study, τ . This is done in the specific case of the "ramp plateau" weighting function which was used for interim monitoring and reporting in the NLST. This is the function which takes the value 0 at $t = 0$, has linear increase to the value 1 at $t = t_c$ and then maintains this constant value forward.

$$Q(t) = \frac{t}{t_c} \wedge 1 \quad (8.4)$$

In the NLST, the value of $t_c = 4$ years was used. Next, by imposing some mild assumptions we will be able to express all quantities in the integrands in terms of the cross-arm pooled cancer mortality cumulative hazard function, H and thereby solve the integrals via a simple change of variables. The resulting expressions require only values of $H(t)$ at $t = t_c$, $t = \tau - t_{er}$ and $t = \tau$, where t_{er} is the calendar time at which randomization was concluded. First we shall list the required assumptions. In the following discussion, S , S_{lr} and S_{oth} are survival functions corresponding to the cross-arm pooled cancer mortality, administrative censoring or "live removal" and other cause mortality. The latter two were the only sources of censoring in the NLST because complete ascertainment with respect to mortality was possibly through the use of the matching death certificates through the national death index.

Condition 8.1. *Other cause mortality is proportional to cancer mortality, i.e. that $\theta = -d\log(S_{oth})/dH$ is constant.*

Condition 8.2. *Proportional allocation: $e(\xi, 0) \equiv e(0, 0)$.*

Condition 8.3. *Accrual is uniform on the scale of H , so that*

$$S_{lr}(\xi) = \frac{H(\tau) - H(\xi)}{H(\tau) - H(\tau - t_{er})} \wedge 1, \quad (8.5)$$

where τ is the time at which the required number of events are obtained, and t_{er} is the time at which randomization is completed.

Condition 8.4.

$$Q(\xi) = \frac{\xi}{t_c} \wedge 1 \equiv \frac{1 - \exp(-H(\xi) \wedge H(t_c))}{1 - \exp(-H(t_c))}. \quad (8.6)$$

The other cause versus cancer proportionality assumption is perhaps the most arguable. However, the extent to which it is violated in practice has little impact upon our results as other cause mortality enters our results only through its survival function which maintains a value in excess of 0.95 throughout the trial. The proportional allocation assumption approximates what we see in practice quite closely, especially in the case of a large trial of a rare event. In the NLST there was 1 to 1 randomization so that $e(0, 0) = 1/2$. The extent to which the latter two assumptions 8.3 and 8.4 hold both depend upon the extent to which pooled cancer specific mortality grows at a constant rate. In the case of the NLST, the pooled cancer mortality cumulative hazard function did grow at an approximately linear rate.

Variance at Planned Termination

$$\begin{aligned} v(\tau) &= \langle Q|F|Q \rangle_\tau = \int_0^\tau Q^2(\xi)e(\xi, 0)(1 - e(\xi, 0)) dG(\xi) \\ &= \int_0^\tau Q^2(\xi)e(\xi, 0)(1 - e(\xi, 0)) S_{oth}(\xi)S_{lr}(\xi)S(\xi)dH(\xi). \quad (8.7) \end{aligned}$$

Here, S , S_{lr} and S_{oth} are survival functions corresponding to the cross-arm pooled cancer mortality, administrative censoring or “live removal” and other cause mortality. The latter two were the only sources of censoring in the NLST because complete ascertainment with respect to mortality was possibly through the use of the matching death certificates through the national death index. Therefore, we can express the differential, dG , in this way. Under assumptions 8.1, 8.2, 8.3, and 8.4, we apply the change of variables, $\eta = H(\xi)$, to obtain

$$\begin{aligned} v(\tau) &= \frac{1}{4} \int_0^{H(\tau)} \left(1 - e^{-(\eta \wedge H(t_c))}\right)^2 e^{-\theta\eta} \left\{ \frac{H(\tau) - \eta}{H(\tau) - H(\tau - t_{er})} \wedge 1 \right\} e^{-\eta} d\eta \\ &= \frac{1}{4} \int_0^{H(t_c) \wedge H(\tau - t_{er})} (1 - 2e^{-\eta} + e^{-2\eta}) e^{-(\theta+1)\eta} d\eta \\ &\quad + \frac{I(t_c < \tau - t_{er})}{4} (1 - e^{-H(t_c)})^2 \int_{H(t_c)}^{H(\tau - t_{er})} e^{-(\theta+1)\eta} d\eta \\ &\quad + \frac{I(\tau - t_{er} < t_c)}{4(H(\tau) - H(\tau - t_{er}))} \int_{H(\tau - t_{er})}^{H(t_c)} (1 - 2e^{-\eta} + e^{-2\eta}) e^{-(\theta+1)\eta} (H(\tau) - \eta) d\eta \\ &\quad + \frac{(1 - e^{-H(t_c)})^2}{4(H(\tau) - H(\tau - t_{er}))} \int_{H(\tau - t_{er}) \vee H(t_c)}^{H(\tau)} e^{-(\theta+1)\eta} (H(\tau) - \eta) d\eta \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned}$$

These evaluate to:

$$\begin{aligned}
I_1 &= \frac{1}{4} \left\{ \frac{1 - e^{-(\theta+1)H_m}}{\theta+1} - 2 \frac{1 - e^{-(\theta+2)H_m}}{\theta+2} + \frac{1 - e^{-(\theta+3)H_m}}{\theta+3} \right\} \text{ where } H_m = H(t_c) \wedge H(\tau - t_{er}), \\
I_2 &= I(t_c < \tau - t_{er}) \left(1 - e^{-H(t_c)} \right)^2 \frac{e^{(\theta+1)H(t_c)} - e^{-(\theta+1)H(\tau - t_{er})}}{4(\theta+1)}, \\
I_3 &= \frac{I(\tau - t_{er} < t_c)}{4(H(\tau) - H(\tau - t_{er}))} \\
&\quad \times \left\{ \left(\frac{e^{-(\theta+1)H(\tau - t_{er})}}{\theta+1} - 2 \frac{e^{-(\theta+2)H(\tau - t_{er})}}{\theta+2} + \frac{e^{-(\theta+3)H(\tau - t_{er})}}{\theta+3} \right) (H(\tau) - H(\tau - t_{er})) \right. \\
&\quad - \left(\frac{e^{-(\theta+1)H(t_c)}}{\theta+1} - 2 \frac{e^{-(\theta+2)H(t_c)}}{\theta+2} + \frac{e^{-(\theta+3)H(t_c)}}{\theta+3} \right) (H(\tau) - H(t_c)) \\
&\quad - \left(\frac{e^{-(\theta+1)H(\tau - t_{er})} - e^{-(\theta+1)H(t_c)}}{(\theta+1)^2} - 2 \frac{e^{-(\theta+2)H(\tau - t_{er})} - e^{-(\theta+2)H(t_c)}}{(\theta+2)^2} \right. \\
&\quad \left. \left. + \frac{e^{-(\theta+3)H(\tau - t_{er})} - e^{-(\theta+3)H(t_c)}}{(\theta+3)^2} \right) \right\}, \\
I_4 &= \frac{(1 - e^{-H(t_c)})^2}{4(\theta+1)} \\
&\quad \times \left\{ \frac{H(\tau) - (H(\tau - t_{er}) \vee H(t_c))}{H(\tau) - H(\tau - t_{er})} e^{-(\theta+1)(H(\tau - t_{er}) \vee H(t_c))} - \frac{e^{-(\theta+1)(H(\tau - t_{er}) \vee H(t_c))} - e^{-(\theta+1)H(\tau)}}{(\theta+1)(H(\tau) - H(\tau - t_{er}))} \right\}
\end{aligned}$$

respectively.

First Moment at Planned Termination

$$\begin{aligned}
m(\tau) &= \int_0^\tau Q(\xi) e(\xi, 0) (1 - e(\xi, 0)) dG(\xi) \\
&= \int_0^\tau Q(\xi) e(\xi, 0) (1 - e(\xi, 0)) S_{oth}(\xi) S_{lr}(\xi) S(\xi) dH(\xi). \quad (8.8)
\end{aligned}$$

Under assumptions 8.1, 8.2, 8.3, and 8.4, we again apply the change of variables,

$\eta = H(\xi)$, to obtain

$$\begin{aligned}
m(\tau) &= \frac{1}{4} \int_0^{H(\tau)} \left(1 - e^{-\eta \wedge H(t_c)}\right) e^{-\theta \eta} \left\{ \frac{H(\tau) - \eta}{H(\tau) - H(\tau - t_{er})} \wedge 1 \right\} e^{-\eta} d\eta \\
&= \frac{1}{4} \int_0^{H(t_c) \wedge H(\tau - t_{er})} (1 - e^{-\eta}) e^{-\theta \eta} e^{-\eta} d\eta \\
&\quad + \frac{1}{4} I(t_c < \tau - t_{er}) \left(1 - e^{-H(t_c)}\right) \int_{H(t_c)}^{H(\tau - t_{er})} e^{-\theta \eta} e^{-\eta} d\eta \\
&\quad + \frac{1}{4} I(t_c > \tau - t_{er}) \int_{H(\tau - t_{er})}^{H(t_c)} (1 - e^{-\eta}) e^{-\theta \eta} \frac{H(\tau) - \eta}{H(\tau) - H(\tau - t_{er})} e^{-\eta} d\eta \\
&\quad + \frac{1}{4} I(t_c < \tau) \left(1 - e^{-H(t_c)}\right) \int_{H(t_c) \vee H(\tau - t_{er})}^{H(\tau)} e^{-\theta \eta} \frac{H(\tau) - \eta}{H(\tau) - H(\tau - t_{er})} e^{-\eta} d\eta \\
&= J_1 + J_2 + J_3 + J_4
\end{aligned}$$

These evaluate to

$$\begin{aligned}
J_1 &= \frac{1}{4} \left\{ \frac{1 - e^{-(\theta+1)(H(t_c) \wedge H(\tau - t_{er}))}}{\theta + 1} - \frac{1 - e^{-(\theta+2)(H(t_c) \wedge H(\tau - t_{er}))}}{\theta + 2} \right\}, \\
J_2 &= \frac{1}{4} I(t_c < \tau - t_{er}) \left(1 - e^{-H(t_c)}\right) \frac{e^{-(\theta+1)H(t_c)} - e^{-(\theta+1)H(\tau - t_{er})}}{\theta + 1}, \\
J_3 &= \frac{I(t_c > \tau - t_{er})}{4(H(\tau) - H(\tau - t_{er}))} \\
&\quad \times \left\{ \left(\frac{(H(\tau) - H(\tau - t_{er})) e^{-(\theta+1)H(\tau - t_{er})} - (H(\tau) - H(t_c)) e^{-(\theta+1)H(t_c)}}{\theta + 1} \right. \right. \\
&\quad \left. \left. - \frac{(H(\tau) - H(\tau - t_{er})) e^{-(\theta+2)H(\tau - t_{er})} - (H(\tau) - H(t_c)) e^{-(\theta+2)H(t_c)}}{\theta + 2} \right) \right. \\
&\quad \left. - \left(\frac{e^{-(\theta+1)H(\tau - t_{er})} - e^{-(\theta+1)H(t_c)}}{(\theta + 1)^2} - \frac{e^{-(\theta+2)H(\tau - t_{er})} - e^{-(\theta+2)H(t_c)}}{(\theta + 2)^2} \right) \right\} \\
J_4 &= \frac{I(t_c < \tau) (1 - e^{-H(t_c)})}{4(H(\tau) - H(\tau - t_{er}))} \\
&\quad \times \left\{ \frac{(H(\tau) - H(t_c \vee (\tau - t_{er}))) e^{-(\theta+1)H(t_c \vee (\tau - t_{er}))}}{\theta + 1} - \frac{e^{-(\theta+1)H(t_c \vee (\tau - t_{er}))} - e^{-(\theta+1)H(\tau)}}{(\theta + 1)^2} \right\}
\end{aligned}$$

respectively.

Duration of Trial

The duration the NLST was part of the design. In other situations in which the design stipulates that the trial should run until required number of events is attained, the above change of variables technique can be used to find a closed form expression for

$$G(\tau) = \int_0^\tau S_{oth}(\xi)S_{lr}(\xi)S(\xi)dH(\xi). \quad (8.9)$$

in terms of the projected values of H at $t = \tau$ and $t = \tau - t_{er}$. Then using the plug-in estimate $\mathbb{E}N_n(\tau)/n$ for $G(\tau)$ this expression can be inverted to solve for τ , the duration of the trial.

8.3. Sampling density of $(J, X_n(t_J))$

As in Armitage, McPherson and Rowe, [1], the sampling density of $(J, X_n(t_J))$ can be derived recursively as follows. Let $\Delta_j = f_{n,j} - f_{n,j-1}$ and let $\phi_v(x) = \phi(x/\sqrt{v})/\sqrt{v}$ where ϕ is the density of the standard normal. First,

$$\pi((1, x)) = \phi_{f_1}(x). \quad (8.10)$$

Next, for all $j > 1$,

$$\begin{aligned} & \pi((j, x) ; \mathbf{b}_{1:(j-1)}, \mathbf{f}_{1:j}) \\ &= \int_{-\infty}^{\sqrt{f_{j-1}b_{j-1}}} \pi((j-1, \xi) ; \mathbf{b}_{1:(j-2)}, \mathbf{f}_{1:(j-1)}) \phi_{\Delta_j}(x - \xi) d\xi \end{aligned} \quad (8.11)$$