Statistics and Its Interface Volume 7 (2014) 27–42

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

Grant Izmirlian

This paper considers a group sequentially monitored trial on a survival endpoint, monitored using a weighted log-rank (WLR) statistic with bounded weight function. Results and discussion center on the theoretical and practical concerns encountered in this setting. We demonstrate the construction of Lan Demets boundaries and calculation of design adjusted p-values and confidence intervals. Central to this discussion is a bijection which exists between the weighted log-rank statistic and a weighted averaged logged relative risk which is shown to exist when there is no sign change in the weighting function. We also show under parametric assumptions that the bijection can be estimated at each interim analysis. The results are benchmarked in a simulation study and a section is devoted to the decisions made in applying the methods to monitoring in the the National Lung Screening Trial (NLST).

AMS 2000 subject classifications: Primary 62L12; secondary 62N022.
Keywords and phrases: Weighted log-rank statistic, Group sequential, Interim analysis, Estimation.

1. INTRODUCTION

Randomized controlled cancer incidence and mortality screening trials, and two armed survival endpoint trials in general, make use of the log-rank statistic in tests of between group differences in rates. When the intervention arm hazard rate is proportional to the control arm hazard, the log-rank test, being the MLE in this case, is most powerful for a single hypothesis test. Even when the proportionality assumption does not hold, the relative risk is still used as the main summary parameter in most if not all of such trials. In a single test of hypothesis, departures from proportional hazards result in large losses in power for hypothesis tests based upon the relative risk. In cancer screening trials, and disease prevention trials in general, typically the intervention is offered annually for, say, three to five years and then subjects are followed up for either the planned duration of the trial or until a planned number of events have occurred. In these cases we expect a delayed benefit, increasing from zero to a maximum value and then a leveling off at a constant value, with attenuation if the follow-up continues beyond the duration of this maximum efficacy. In such cases when a good guess at the true shape of the logged hazard ratio exists, the weighted log-rank (WLR) statistic with weights approximately proportional to the true shape have optimal power for a single test of hypothesis. For this reason a WLR statistic was used in monitoring and reporting in the Women's Health Initiative [1]. In cancer screening trials, we expect the benefit to nadir at the last offered screen plus the median lead time associated with the particular screening modality. Therefore, if follow-up is concluded before the effectiveness of screening becomes attenuated, a reasonable guess at the magnitude of the instantaneous logged relative risk is linear growth from zero to this nominal expected time of maximum benefit and then flat thereafter, so that a weight function based upon this shape should demonstrate gains in power. Besides gains in power, a deterministic weighting function of this type has other desirable characteristics [5]. We used this type of WLR statistic in the monitoring and reporting of the National Lung Screening Trial (NLST) [11, 10]. We mention in passing that there are other models allowing for a time varying hazard ratio in the literature. Wieand et al. suggests the use of the integrated difference between survival functions, while Yang and Prentice presents a statistic that uses the control arm survival function to switch between an early and a late effect, while Nan et al. approaches the problem using piecewise constant proportional hazards.

Institutional Review Board regulatory process requires the monitoring of clinical trials via interim analysis of the primary endpoint at regularly scheduled meetings of a data safety and monitoring board. The machinery of sequential design allows for multiple hypothesis tests based upon the sequentially maturing data. Rather than Bonferroni adjustment, the total probability type I error is allocated unequally so that less is "spent" in earlier analyses based upon immature data, reserving the bulk of it for the final analysis. This allocation of type I error probability is often done using the spending function approach of Lan and DeMets. Because the analysis number at which the trial stops now

arXiv: 1102.5088

joins the test statistic to form a bivariate observation, the sequential design fundamentally changes the nature of the inferential scheme. There is no longer a most powerful test of hypothesis under any given model, we must reconsider the notion of order in deriving tail probabilities, and p-values under the null hypothesis are no longer uniformly distributed. For these reasons, the planning of a sequential design or interim analysis plan requires simulation under a variety of possible trial scenarios in which candidate designs consisting of boundary construction method and test statistic choice can be benchmarked. The design of the NLST stipulated the use of a WLR statistic for the reasons mentioned above. The methodology discussed here resulted from our study of the usual machinery as it exists in the case of the unweighted log-rank statistic and our need to adapt it to the setting of the weighted log-rank statistic. The method whereby the WLR statistic can be rescaled and interpreted as a weighted averaged logged relative risk arose out of the need to include a clinically meaningful summary in the main report on the NLST, combined with the desire to have concurrence between the statistic as it is used to monitor the trial and the statistic presented in the main report.

The aim of this paper is to point out how the weighting of events affects the statistical monitoring the trial vis a vis the information fraction, the construction of a futility boundary and estimation of parameters and confidence interval when the trial is stopped. Central to this discussion will be the connection between hypothesis testing on the standard normal scale and prediction on the more clinical meaningful scale of relative risk. We will show that as is the case with the un-weighted log-rank statistic, the asymptotic distribution of the WLR statistic, suitably normalized is a mean zero Gaussian process plus a time varying non-linear drift. As is the case with the unweighted log-rank statistic, the drift function has value at the planned conclusion equal to the square root of the variance function times the weighted average logged relative risk. In the following subsection we will investigate the form of the drift function at interim analyses. We will see that it is equal to the square root of the variance at the planned conclusion times the information fraction times a dynamic correction factor related to the inner product between the chosen weighting function and the true shape function.

2. TERMINOLOGY AND FRAMEWORK

2.1 The main result and its corollary

In this section we must introduce a minimum amount of notation necessary to express the weighted log-rank statistic, asymptotically, as a mean zero Gaussian process plus a drift function. The setting is 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 Z_i indicate membership in the intervention arm ($Z_i = 1$) or control arm ($Z_i = 0$). We assume, conditional upon Z_i, that individuals, $i = 1, \ldots, n$ are distributed independently and identically. Let $dH_0(t)$ and $dH_1(t)$ be the trial arm specific increments in cumulative hazard functions. In order that an instantaneous hazard ratio is not undefined, we must assume that $H_1(t)$ is absolutely continuous with respect to $H_0(t)$:

Condition 2.1. *$dH_1(t)$ is absolutely continuous with respect to $dH_0(t)$*

Now we can write the instantaneous logged hazard ratio, as the log of the Radon-Nikodym derivative:

$$\beta(t) = \log\left\{\frac{dH_1(t)}{dH_0(t)}\right\}. \tag{1}$$

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. We will use the function, $Q : [0, \tau] \to \mathbb{R}_+$, to weight events occurring at time t.

The $\sqrt{n}$-normalized log-rank statistic with weighting function Q at follow-up time t is:

$$U_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t Q(\xi) \{Z_i - E_n(\xi, 0)\} dN_i(\xi). \tag{2}$$

Its estimated variance is:

$$V_n(t) = \frac{1}{n} \int_0^t Q^2(\xi) E_n(\xi, 0) (1 - E_n(\xi, 0)) dN_n(\xi) \tag{3}$$

In the above expressions, $E_n(t, 0)$ represents the proportion of the population at risk at time t that is in the intervention arm:

$$\begin{aligned} R_n(\xi, 0) &= \frac{1}{n} \sum_{i=1}^{n} I(T_i \geq \xi) \\ \text{and} & \\ E_n(\xi, 0) &= \frac{1}{nR_n(\xi, 0)} \sum_{i=1}^{n} Z_i I(T_i \geq \xi). \end{aligned}$$

The meaning of the second argument and why it is set to zero in the above will be made clear below. We remark in passing that when the primary outcome is very rare relative to the initial sample size and censoring is balanced between the arms, which is of course the case in cancer mortality trials then $E_n(t, 0)$ is nearly identically equal $1/2$ throughout the duration of the trial. This is stated formally as an assumption in a later section devoted to calculation of relevant quantities at interim analyses, when (and if) the trial is stopped early, or if it continues on to the planned closeout. By lemma 7.1 listed in appendix 7 it follows that $R_n(t, 0)$

and $E_n(t,0)$ have almost sure limits, $R(t,0)$ and $e(t,0)$, respectively.

Note that we have not yet posed any assumptions on the true form of the instantaneous logged relative hazards ratio. Under the assumption that the true shape is proportional to the chosen weighting function then U_n is the score process associated with the logged partial likelihood. We discuss this and another possible shape assumption immediately following the statement of the main asymptotics result, but unless otherwise stated, we do not restrict the discussion to any such shape assumption. Our next task is to express U_n, asymptotically, as the sum of a mean zero Gaussian process plus a drift function. Towards this end we defined the weighed average logged relative risk summary of the instantaneous logged relative risk and view it as a kind of dot product between the chosen weighting function and the true shape function.

First, some more necessary notation. Let $G_n(t) = N_n(\xi)/n$ and let $\mathbb{F}_n(t) = \int_0^t E_n(\xi,0)(1 - E_n(\xi,0))\, dG_n(t)$. Note that $\mathbb{F}_n$ is the variance of the unweighted log-rank statistic and so naturally, the variance of the weighted log-rank statistic is expressed as an integral against its increments. By lemma 7.1 of appendix 7 it follows that $G_n(t)$ and $\mathbb{F}_n(t)$ have almost sure limits, G and $\mathbb{F}$, pointwise.

Next, we introduce the following notation for cross moment integrals against $d\mathbb{F}$ over $(0,t)$:

$$\langle \psi_1 | \mathbb{F} | \psi_2 \rangle_t = \int_0^t \psi_1(\xi)\, \psi_2(\xi) d\mathbb{F}(\xi)\,. \tag{4}$$

We can now define the following summary of the instantaneous logged relative risk function—its weighted average against the measure $Q\, d\mathbb{F}$:

$$\beta^\star = \frac{\langle Q|\mathbb{F}|\beta\rangle_\tau}{\langle Q|\mathbb{F}|1\rangle_\tau}\,. \tag{5}$$

Having defined the weighted average logged relative risk, we now have representation of the instantaneous logged relative risk function, $\beta(t) = \beta^\star\, q(t)$ as the product of its weighted average value, $\beta^\star$ times a shape function, $q = \beta(t)/\beta^\star$. Note that it follows that the shape function has weighted average value equal to 1:

$$1 = \frac{\langle Q|\mathbb{F}|q\rangle_\tau}{\langle Q|\mathbb{F}|1\rangle_\tau}\,. \tag{6}$$

We also remark in passing that the estimated variance, V_n, of the weighted log-rank statistic, U_n, given above is a cross moment: $V_n(t) = \langle Q|\mathbb{F}_n|Q\rangle_t$. The variance, V_n, is the second moment of Q with respect to $\mathbb{F}_n$. Its first moment will also enter the discussion. Put $m_n(t) = \langle Q|\mathbb{F}_n|1\rangle_t$. The variance ratio or information fraction is $f_n(t;\tau) = V_n(t)/V_n(\tau)$. The first moment fraction will also enter the discussion. Put $r_n(t;\tau) = m_n(t)/m_n(\tau)$. Pointwise almost sure convergence of these quantities to their limits, $v(t), m(t), f(t;\tau)$, and $r(t;\tau)$, respectively, is a consequence of lemma 7.1 of appendix 7.

We are now in a position to express the weighted log-rank statistic, asymptotically, as a mean zero Gaussian process plus a drift function. In the theorem that follows, we consider the weighted log-rank (WLR) statistic at time t on the "Brownian scale": $X_n(t) = U_n(t)/\sqrt{V_n(\tau)}$. On occasion we also consider the WLR statistic on the standard normal scale $Y_n(t) = U_n(t)/\sqrt{V_n(t)}$. In addition to the absolute continuity assumption given above, we need two boundedness assumptions required in the proofs of lemma 7.1 of appendix 7.

Condition 2.2. *The control arm cumulative hazard function, $H_0(t)$ and the chosen weighting function, Q, are bounded on $[0,\tau]$.*

Condition 2.3. *The instantaneous logged relative risk function, $\beta(t)$ is bounded on $[0,\tau]$.*

Theorem 2.1. *Under conditions 2.1, 2.2 and 2.3, and under the family of local alternatives, $\beta_n^\star = b^\star/\sqrt{n}$, the score statistic, normalized to the "Brownian scale" is asymptotically a Brownian motion on $[0,1]$ plus a drift.*

$$X_n(t) \xrightarrow{\mathcal{D}} W(f(t;\tau)) + \mu(t) \tag{7}$$

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

$$\mu(t) = \frac{\langle Q|\mathbb{F}|q\rangle_t}{\sqrt{\langle Q|\mathbb{F}|Q\rangle_\tau}}\, b^\star\,. \tag{8}$$

The proof of Theorem 2.1 is given in appendix 7.

2.2 Rational for weighting function and departures from initial guesses

As mentioned in the introduction, we have in the setting of cancer mortality screening trials an expectation of delayed benefit. In this case, a good guess at the form of the true shape function, q, is a linear rise from a value of zero at time zero to its maximum value at some time t_q and then level thereafter:

$$\mathrm{Ramp}[t_q](t) = \frac{t}{t_q} \wedge 1 \tag{9}$$

which we can call the "ramp-plateau" function. We will suppress dependence on t on occasion and write $Ramp[t_q]$ for the ramp-plateau function reaching its maximum value at $t = t_q$. If our trial is stopped before efficacy begins to decline, then we can be fairly certain that the true shape is of this form. Let t_q be the true time to maximum benefit. If monitoring the trial were not the central goal, and we started with complete data at the end of the trial, we could treat t_q as a parameter in the model and estimate its value.

The need to monitor the trial beginning with times possibly before t_q complicates the matter and causes a situation in which t_q as a parameter would be unidentifiable at analysis times prior to it. Therefore, we must guess at the value of t_q. In the introduction, we proposed setting the nominal (guessed) time to maximum efficacy to the time on study at the last offered screen plus the median lead time. In the following, we denote by t_Q the nominal time to maximum efficacy which we will use in the definition of the weighting function as the time at which it flattens. If our guess, t_Q, is perfect and $t_Q = t_q$ then our WLR statistic has optimal power for a single analysis with complete data. However, it is well known that for a sequential analysis there is no optimal test statistic. We will return to these considerations in a later section devoted to numerical study.

At this point we focus discussion on the drift function, its relationship to the weighted logged relative risk, $\beta^\star$ and the consequences of various shape assumptions or the absence of any shape assumptions. First we rewrite the drift function in the following form:

$$\mu(t) = \frac{\langle Q|\mathbb{F}|1\rangle_\tau}{\sqrt{\langle Q|\mathbb{F}|Q\rangle_\tau}}\,\rho(t;\tau) f(t;\tau)\sqrt{n}\beta^\star\,, \tag{10}$$

where

$$\rho(t;\tau) = \frac{\langle Q|\mathbb{F}|q\rangle_t / \langle Q|\mathbb{F}|1\rangle_\tau}{\langle Q|\mathbb{F}|Q\rangle_t / \langle Q|\mathbb{F}|Q\rangle_\tau}\,. \tag{11}$$

When the true shape is proportional to the chosen weighting function, then $\rho \equiv 1$, and the drift is linear in the information fraction as is the case in the unweighted log-rank statistic under proportional hazards. Note that in general, without any shape assumptions on the instantaneous hazard ratio, q, the drift function is equal to this $\rho \equiv 1$ version times the time dependent correction factor, ρ. By equations 6 and 8, it follows that $\rho(\tau;\tau) = 1$ so that the value of the drift function at the scheduled end of the trial is

$$\mu(\tau) = \frac{\langle Q|\mathbb{F}|1\rangle_\tau}{\sqrt{\langle Q|\mathbb{F}|Q\rangle_\tau}}\,b^\star\,, \tag{12}$$

again, without any assumptions on the shape of the instantaneous hazard ratio, q.

With regard to estimation of $\beta^\star$ we have, even in this general set of circumstances, the following corollary:

Corollary 2.1. *At the planned conclusion of the trial, τ, an estimate of $\beta^\star$ is given by the following:*

$$\widehat{\beta}^\star = X_n(\tau)\frac{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}{\sqrt{n}\,\langle Q|\mathbb{F}_n|1\rangle_\tau}\,. \tag{13}$$

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

$$\mathrm{var}\left[\widehat{\beta}^\star\right] = \frac{\langle Q|\mathbb{F}_n|Q\rangle_\tau}{n\,\langle Q|\mathbb{F}_n|1\rangle_\tau^2}\,. \tag{14}$$

Table 1. Two possible shape assumptions for the shape of the instantaneous hazard ratio, q, and the resulting form of the correction factor, ρ

Assumed form of q	ρ
$q \propto Q$	$\rho \equiv 1$
$q \equiv 1$	$\rho = r(t)/f(t)$

Notice that when $Q \equiv 1$ and X_n is the unweighted log-rank statistic, the estimate for $\beta^\star$ in expression 13 reduces to a more familiar form since the ratio of inner products cancels leaving $\sqrt{V_n(\tau)}$.

2.3 Estimates of $\beta^\star$ in a trial stopped early

At the planned conclusion of the trial, expression 13 provides an estimate for $\beta^\star$ in terms the WLR statistic, its variance and one other functional of the chosen weighting function, without any required assumptions of the form of the true shape function, q. However, when the trial is stopped early, the unknown true shape function, q, remains in the expression for the drift function so that in order to obtain an estimate for $\beta^\star$ we must impose additional assumptions on the form of the true shape. Specifically, we have the following estimate for $\beta^\star$ at an early conclusion at time a $t_J < \tau$:

$$\widehat{\beta}^\star = \frac{X_n(t_J)}{f_n(t_J)}\frac{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}{\sqrt{n}\langle Q|\mathbb{F}_n|1\rangle_\tau\,\rho_n(t_J)}\,, \tag{15}$$

and the following estimated variance:

$$\mathrm{var}\left[\widehat{\beta}^\star\right] = \frac{\langle Q|\mathbb{F}_n|Q\rangle_\tau}{n\,f_n(t_J;\tau)\,\langle Q|\mathbb{F}_n|1\rangle_\tau^2\,\rho_n^2(t_J)}\,. \tag{16}$$

Here, ρ_n, is the correction factor which requires complete knowledge of the shape function, q. In order to estimate $\beta^\star$ in a trial stopped early, we must place additional assumptions on q. At a minimum in order to have a monotone drift function which is necessary for proper monitoring, we require the following.

Condition 2.4. *The shape function, q, is non-negative.*

Estimation of $\beta^\star$ at a trial stopped early requires more than the assumption of non-negativity in q. In table 1 we consider two possible forms of the true shape function, q, each resulting from an additional semi-parametric assumption and the resulting form of the correction factor, ρ. In the first row of table 1, consider the possibility that the true shape, is proportional to the chosen weighting function, $q = K\,Q$. In this case, the correction factor, ρ, is identically 1. This is because there is cancellation in the ratio of inner products to time t in the numerator of expression 11, and the ratio of inner products to time τ in the denominator of expression 11 reduces to the constant of proportionality, K, by the unitary property of q, shown in expression 6. In this case the drift function is linear in the information fraction,

f, resulting in the usual estimate of the logged relative risk under proportional hazards when the unweighted log-rank statistic is used.

In the second row of table 1, consider the possibility that the proportional hazards assumption is true, and the true shape function is identically 1, but as this was not anticipated, we are using a WLR statistic with some non-constant weighting function, Q. In this case, the correction factor, ρ is the ratio of the Q-first moment fraction, $r(t;\tau) = \langle Q|\mathbb{F}_n|1\rangle_t / \langle Q|\mathbb{F}_n|1\rangle_\tau$, to the usual variance fraction, which is the time scale onto which the Gaussian process is transformed, $f(t) = \langle Q|\mathbb{F}_n|Q\rangle_t / \langle Q|\mathbb{F}_n|Q\rangle_\tau$. In this case there is cancellation in the product of f and ρ in the denominator of expression 11 leaving the new information scale: $f(t)\,\rho(t) = r(t)$ so that the drift is linear in the Q-first moment fraction, r, which is distinct from the time scale on which the statistic normalized to a Gaussian process. Since the use of the WLR statistic is based upon belief that the true shape is a ramp function of some kind, then the best approach would be to stipulate the $q = KQ$ assumption as part of the interim analysis plan. Results based upon the $q \equiv 1$ assumption can be presented in a secondary analysis.

To summarize, under the $q \propto Q$ assumption, we have

$$\widehat{\beta}^\star = \frac{X_n(t_J)}{f_n(t_J)} \frac{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}{\sqrt{n}\langle Q|\mathbb{F}_n|1\rangle_\tau}, \tag{17}$$

with estimated variance,

$$\text{var}\left[\widehat{\beta}^\star\right] = \frac{\langle Q|\mathbb{F}_n|Q\rangle_\tau}{n\, f_n(t_J;\tau)\, \langle Q|\mathbb{F}_n|1\rangle_\tau^2}. \tag{18}$$

When we assume that $q \equiv 1$ even though we are using non-constant weights, we have

$$\widehat{\beta}^\star = \frac{X_n(t_J)}{r_n(t_J)} \frac{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}{\sqrt{n}\langle Q|\mathbb{F}_n|1\rangle_\tau}, \tag{19}$$

with estimated variance,

$$\text{var}\left[\widehat{\beta}^\star\right] = \frac{\langle Q|\mathbb{F}_n|Q\rangle_\tau\, f_n(t_J;\tau)}{n\, r_n(t_J;\tau)\, \langle Q|\mathbb{F}_n|1\rangle_\tau^2\, r_n(t_J;\tau)}. \tag{20}$$

In the next section we outline how some of the typical steps in the design and execution of a monitoring plan must be modified to accommodate the use of a WLR statistic, with particular focus on choices made in the design and execution of the monitoring plans for the NLST cancer screening trial.

3. APPLICATION TO MONITORING AND FINAL REPORTING IN A CLINICAL TRIAL

3.1 Design of monitoring plan

Institutional review board approval requires the stipulation of a statistical design which clearly states the main outcome measure, intervention(s), hypothesis, and statistical power to test the hypothesis. The stipulation of a monitoring plan for early termination due to possible harm or overwhelming evidence of benefit is also a key requirement but often the technical details of the monitoring plan are not spelled out. Good practice dictates that a monitoring plan should be stipulated with the initial design and if not at that point, then soon after, and certainly prior to any presentation of data to a data safety and monitoring board (DSMB). Design of a monitoring plan is done by considering possible trial scenarios, e.g. the range of the occurrence rate of the event in the control arm, the range of the efficacy parameter, which should encompass both efficacy and harm, and a list of candidate monitoring plans given by possible choices for the test statistic, the boundary construction method and the timing of analyses. In the monitoring of the NLST trial, we used the Lan-Demets procedure, [8], to construct efficacy boundary points, and to separately construct non-binding futility boundary points. We remark here that, following consensus, we recommend using a non-binding futility boundary which is constructed after the construction of an efficacy boundary. This is preferred to joint construction of efficacy and futility boundaries as that approach results in a discounted efficacy criterion. One then conducts a simulation study in which each trial scenario is used to simulate replicate trials, and for each replicate trial, each candidate monitoring plan is applied to the trial in simulated trial time to monitor the data at the planned analysis times. In this manner operating characteristics such as the average power over efficacious scenarios and the duration of the trial under harmful scenarios can be determined. There are a variety of open source tools available for this type of investigation in the design of a monitoring plan. One which incorporates the necessary modifications to the theory to allow for the use of the weighted log rank statistic is "PwrGSD", [6]. The interested reader should refer to the included package vignettes, [7].

3.2 Construction of boundary points during run of trial

Both efficacy and futility boundary points require the variance information fraction at each analysis. For this purpose, a value of the end of trial variance is required. If censoring patterns are identical in the two trial arms as they should be and we have "equal allocation" which means a consistent 1 to 1 balance in the risk sets throughout run of the trial, then in settings in which an unweighed log-rank statistic is used, the end of trial variance is just 1/4 of the end of trial events, which is a design stipulated quantity. When a WLR statistic is used, then ideally an end of trial variance should have been included in the design of the interim analysis plan, having been based upon simulation study as mentioned above, and then the trial can run until this value is attained, e.g. a maximum information design. Often times a maximum funding design is imposed upon the

investigators after the fact. In such cases a predicted value is required. We show in appendix 8.1 how this may be done. If futility boundary points are desired, then values of the drift function are also required. Values of the drift function at each analysis are easily estimated from the available data using 10, applying either the $q \propto Q$ or the $q \equiv 1$ assumption, and using the design stipulated value of $\beta^\star$ and current values of the variance and Q-first moment functions. If the futility analysis is being done using stochastic curtailment, the value of the drift function at the end of the trial is also required. This additional requirement, of course, makes the use of stochastic curtailment for futility analysis less attractive. If stochastic curtailment is being used and a value of the end of trial drift is required, then expression 12 can be used with $\sqrt{n}\beta^\star$ replacing $b^\star$ in that expression. This requires, in addition to the design stipulated value of $\beta^\star$, values of the end of trial variance and end of trial Q-first moment. These are either design stipulated, as mention above, or predicted using the approach outlined in section 8.1.

3.3 Sampling density

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, $(X_n(t_J), J)$, where J and $X_n(t_J)$ are the analysis number and the value of the weighted log-rank statistic at an efficacy crossing. The sampling density of $(X_n(t_J), J)$ is of the form $p((x, j)$ which we define in terms of g_j, which is in turn, defined recursively. As in Armitage, McPherson and Rowe, if we let $f_1(x) = \phi(x/\sqrt{f_1})/\sqrt{f_1}$ for all x, then $p((x, 1)) = g_1(x)$ for $|x| \geq \sqrt{f_1} b_1$ and $p((x, 1)) = 0$ otherwise. Here, ϕ is the density of the standard normal random variable. For $j > 1$, if we let

$$g_j((x)) = \frac{1}{\sqrt{2\pi(f_j - f_{j-1})}} \times \int_{|x|<\sqrt{f_1} b_1} \phi\left(\frac{x-y}{\sqrt{f_j - f_{j-1}}}\right) g_j(y)\, dy\,, \tag{21}$$

then $p((x, j)) = g_j(x)$ for $|x| \geq \sqrt{f_j} b_j$ and zero otherwise.

Let $\bar{\Pi}((x, j))$ be the probability of the upper tail under $p((x, j))$. 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 $(X_n(t_J), J)$. Here we prefer to use the stage-wise ordering: $(x, j) > (y, k)$ if and only if $(j = k$ 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 positive values of the parameter correspond to efficacy as it is a simple matter to modify these results to the case where negative values of the parameter correspond to efficacy (i.e. the present setting, relative risk).

3.4 P-value

Under the ordering given above, the region further away from the null than $(X_n(t_J), J)$ 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:

$$\bar{\Pi}((X_n(t_J), J)) + \sum_{\ell=1}^{J-1} \bar{\Pi}((\sqrt{f_\ell}\, b_\ell, \ell))) \tag{22}$$

3.5 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 $\widehat{\beta}^\star$ is derived as follows. If we denote by x_u the solution in x of the equation

$$\alpha_{tot} - \alpha_{J-1} = \bar{\Pi}((x, J)) + \sum_{\ell=1}^{J-1} \bar{\Pi}((\sqrt{f_\ell}\, b_\ell, \ell))\,, \tag{23}$$

then the design-adjusted confidence interval is

$$\widehat{\beta}^\star \pm \frac{x_u}{\sqrt{f_{n,J}}} \sqrt{\text{mse}\left[\widehat{\beta}^\star\right]}\,, \tag{24}$$

where $[\widehat{\beta}^\star]$ is the estimated mean-squared error of $\widehat{\beta}^\star$ is the estimate obtained using expression 18 or 20 depending upon the shape assumption. 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.

3.6 Bias adjustment

As mentioned earlier, a naive estimate for $\beta^\star$ in a trial stopped early for efficacy such as the estimator given in expression 18 or 20 is biased away from the null. A technique for removing this bias is based upon the sufficiency of $S = (X_{t_J}, J)$, for the sequence of monitoring statistics on the Brownian scale. The sufficiency of S requires that the drift be linear in the information fraction. As this is true only under the $q \propto Q$ assumption, then bias adjustment can only be done in this case. In this section we assume that we are operating under the $q \propto Q$ assumption, using expression 18. Although there is no minimum variance unbiased estimator of a target parameter in a sequential design, application of the Rao-Blackwell theorem, e.g. taking the conditional expectation given a sufficient statistic, of any unbiased statistic, does preserve the unbiasedness feature. Emerson, [3], noted this and computed the conditional expectation of $X_n(t_1)/f_1$ given S. This is the approach taken in the current setting, whereby $\mathbb{E}\{X_n(t_1)/f_1|S\}$ replaces $X_n(t_J)/f_J$ in expression 18:

$$\hat{\beta}^\star = \mathbb{E}\left\{ \frac{X_n(t_1)}{f_1} \middle| S \right\} \frac{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}{\sqrt{n}\langle Q|\mathbb{F}_n|1\rangle_\tau}\,. \tag{25}$$

Table 2. Power under a beneficial scenario

True Shape	Analysis at year(s)	Ramp[3]	Ramp[4]	Ramp[5]	Ramp[6]	unwtd
Ramp[4]	7	0.896	0.907	0.904	0.894	0.821
Prop. Haz.	7	0.748	0.728	0.712	0.702	0.831
Ramp[4]	4,5,6,7	0.878	0.899	0.898	0.896	0.721
Prop. Haz.	4,5,6,7	0.736	0.723	0.702	0.694	0.813

As shown in [3], one uses Bayes theorem to obtain the density of $(X_n(t_1), 1)$ given S. Note that the density of X_1 has support inside the continuation region for situations in which the trial stops at the second or later analysis.

4. NUMERICAL STUDY

In the previous section, we outlined how the use of the WLR statistic changes the way in which a sequential trial is monitored, from the design of the monitoring plan, construction of boundary points, estimation of the weighted average logged relative risk parameter, bias adjustment, construction of design adjusted confidence interval and p-value. In this section we investigate performance of some of these guidelines via numerical study. In the following sub-sections, we investigate the power of the interim analysis design under a trial scenario in which the intervention results in benefit, the duration of the trial under a trial scenario in which the intervention leads to harm, the performance of the Rao-Blackwell bias adjustment technique, and accuracy of the end of trial variance and Q-first moment predicted using the change of variables formula presented in appendix 8. Our investigation centered around three hypothetical trial scenarios based loosely on parameters used by the author in the design of the NLST interim analysis plan. All hypothetical trial scenarios had a balanced randomization of 50,000 in two years and a control arm cancer mortality of 450 per 100,000 person years—roughly the lung cancer mortality among persons with a 30+ pack-year history of smoking. The three scenarios differed in the specification of the relative risk function. The first scenario had a maximum benefit, a logged relative risk of -0.33, at 4 years (ramped benefit), the second scenario had a constant logged relative risk of -0.20 (flat benefit) and the third scenario had a constant logged relative risk of 0.20 (flat harm). All three scenarios included non-compliance in the form of drop-out from the intervention arm as well as cross-over between the two arms. One thousand simulation replicates were generated in each of the three scenarios. Non-compliance resulted in mean $\beta^\star$ values of -0.20, -0.15 and 0.15 in the ramped benefit, flat benefit and flat harm scenarios, respectively. The methods presented here were used to monitor each simulation replicate trial with analyses occurring at years 3, 4, 5, 6, and 7. Less than 10% of the simulated trials stopped at the first analysis and consequently statistics are presented for stopping times 4, 5, 6, and 7. Trials were monitored for efficacy via a one-sided Lan-Demets boundary with O'Brien-Fleming spending of the probability of type I error, with total probability of type I error set to 0.05. Futility was monitored via a one sided non-binding Lan-Demets boundary with O'Brien-Fleming spending of the probability of type II error, with total probability of type II error set to 0.10 and alternate hypothesis $\beta^\star = \log(0.85)$. The power, duration, end of trial functionals and bias reduced estimate of $\beta^\star$ were examined for monitoring using WLR statistics with weighting Ramp[3], Ramp[4], Ramp[5] or Ramp[6] as well as monitoring using the unweighted log-rank statistic. These calculations were done using the author's package, PwrGSD, [6], for the statistical computing environment, R, [12].

4.1 Power and expected duration

In table 2 we list the power, under a beneficial intervention trial scenario, of the different monitoring statistics when the true shape is ramped versus when the true shape is flat. We list the power under a beneficial trial scenario for WLR statistics having the "ramp" type weighting with plateau beginning in year 3, 4, 5, and 6, as well as for the unweighted log-rank statistic when the true shape is either of the "ramp" type, with maximum benefit realized in year 4, or constant. Power of each of these statistics was calculated under both the Ramp(4) true shape and flat true shape for both a design with a single analysis at year 7, and a design with analyses at years 4, 5, 6 and 7. In the first two lines of table 2, we list power under the single analysis design. As we expect from results concerning optimality of weights proportional to true shape in a single analysis design, we see that when the true shape is Ramp[4], then the Ramp[4] WLR statistic has the largest power, but surprisingly, all of the Ramp weighted statistics have power within 1% of the optimum. The unweighted log-rank statistic shows a loss of nearly 8% power dropping from the optimal power of 90.7% to 82.1%. When the true shape is flat, the unweighted log-rank statistic has optimal power, 83.1%, for the single analysis design has, as expected. In this case the Ramp weighted statistics demonstrate loss of power between 9% to 13%. These results point to the following conclusions. Certainly if the correct shape assumption is made and used to form the basis of a weighting function, or the true shape function is flat and an unweighted statistic is used, the power will be optimal. More importantly, if the true shape is of the form Ramp, then the gain in power over use of an unweighted log-rank statistic is realized even if one makes a relatively bad guess at the peak time. Alternately, if the true shape function is

Table 3. Distribution of trial duration under a harmful scenario. Logged relative risk is 0.20, true shape is Ramp[4] or flat, under various monitoring statistics

True Shape	Duration	Ramp[3]	Ramp[4]	Ramp[5]	Ramp[6]	unwtd
Ramp[4]	4	0.607	0.360	0.209	0.152	0.905
	5	0.381	0.614	0.748	0.793	0.090
	6	0.012	0.026	0.043	0.055	0.005
Prop. Haz.	4	0.733	0.510	0.325	0.258	0.989
	5	0.250	0.458	0.610	0.656	0.010
	6	0.017	0.030	0.061	0.081	0.001
	7	0.000	0.002	0.004	0.005	0.000

flat, then the use of a WLR statistic results in slightly less power loss than would be the case if the unweighted log-rank statistic was used and the true shape was of the Ramp form. Overall, the use of the WLR statistic in the single analysis design results in a slightly lower cost in terms of potential lost power even given that one has to guess at the peak time, t_q. The gains of the WLR statistic are even more dramatic in the multiple analyses design, with more than 82% power for each of the Ramp statistics having peak at 3, 4, 5, or 6 years, under the true shape of Ramp[4] with only 66.3% power for the unweighted log-rank statistic. When the true shape function is flat, the unweighted statistic has 81.2% power, with the WLR statistics losing between 9% and 14% power for the Ramp weightings attaining peak values at 3, 4, 5 and 6 years respectively. Thus even with a bad guess for the peak time, t_q, the cost of using a WLR statistic in terms of power is much less than the cost of using the unweighted log-rank statistic.

In table 3 we list the distribution of the number of analyses required to stop the trial (at the futility boundary in a one sided analysis of efficacy) when the intervention results in harm, for the case that true shape is ramped and the case when the true shape is flat, for simulated trials monitored via each of the different statistics. The first three lines of table 3 show the distribution of stopping times for the harmful scenario when the true shape is of the Ramp[4] form, under each of the candidate monitoring statistics: Ramp with maximum attained at 3, 4, 5, or 6 years, as well as for the unweighted log-rank statistic. The second 4 lines show the distribution of stopping time under harm of each of the statistics when the true shape is flat. It appears that the unweighted log-rank statistic fairs the best, both when the true shape is Ramp and when it is flat, with the probability of termination at the first analysis roughly 92% and 98% when the true shape function is of Ramp form and when it is flat, respectively. The weighted statistics, on the other hand, require two analyses to have probability in excess of 95% to stop under either of the two forms, ramped or flat, of the true shape function.

4.2 End of trial functionals

We showed in appendix 8 how to predict end of trial functions, such as the first and second (variance) Q-moments based upon a simple change of variables made possible by imposing several more or less reasonable assumptions. Since the variance formula for the WLR statistic is the same regardless of whether the true shape is ramped or flat, we consider the predicted end of trial variance of the various weighted statistics via simulations when the true shape is ramped under the efficacious trial scenario. All of the other settings are as described in the previous subsection. In table 4 we list projected and actual end of trial variances for each of the weighted statistics. For the most part the predicted values are within an acceptable range of the actual value, with lower quartiles of error in predicted variance less than 1% attenuated in magnitude and upper quartiles of error less than 10% in magnitude, for all simulated trials stopping early at durations of 5, 6 and 7 years monitored via each of the Ramp[3], Ramp(4) and Ramp[5] weighted statistics at simulated trials stopping early at durations of 4, 5 and 6 years. The performance of the projected end of trial variance in the case of the Ramp[6] weighted statistic is less acceptable, with the lower quartile of error -72% and upper quartile of error -62%. The lower and upper quartiles in percent errors of projected values of end of trial first Q-moment range from roughly 10% to about 50%.

4.3 Raw and bias adjusted estimates of $\beta^\star$

In table 6 we list, in each consecutive pair of lines, simulated mean and 95% confidence interval for raw and bias adjusted estimates, respectively, of $\beta^\star$ derived in each simulated trial. Results are shown according to the stopping time of the trial and by the method used for the end of trial variance method, the form of the true shape, and the type of monitoring statistic. End of trial variances used were either the true value (T) or the predicted value (P). The true shape was either ramped benefit (Ramp[4]) or flat benefit (Flat). The monitoring statistic used was a WLR statistic with weighting Ramp[3], Ramp[4], Ramp[5], or Ramp[6]. As we mentioned in the introductory remarks to this section, in the efficacious scenarious when the shape of the true logged relative function is of ramped form, we used a true value of $\beta^\star$ equal to -0.20. In the efficacious scenarious when the shape of the true logged relative risk function is constant, we used a true value of $\beta^\star$ equal to -0.15. The results in the second 8 lines of table 6 should be compared to this value.

Table 4. Predicted values of end of trial variance, based upon a change of variables technique, and corresponding actual values from simulation

Statistic		Projected End of Trial Var (95% CI)	Actual End of Trial Var (95% CI)
WLR[R(3)]	dur5	0.0036 (0.00323, 0.00403)	0.00344 (0.00325, 0.00363)
	dur6	0.00357 (0.00318, 0.00399)	0.00345 (0.00324, 0.00364)
	dur7	0.00356 (0.00307, 0.00401)	0.00344 (0.00324, 0.00363)
WLR[R(4)]	dur5	0.00302 (0.00269, 0.00343)	0.00282 (0.00266, 0.00298)
	dur6	0.00297 (0.00263, 0.00331)	0.00283 (0.00266, 0.00299)
	dur7	0.00297 (0.00257, 0.00333)	0.00282 (0.00265, 0.00299)
WLR[R(5)]	dur5	0.00235 (0.00213, 0.00258)	0.00223 (0.00209, 0.00236)
	dur6	0.00232 (0.00211, 0.00253)	0.00224 (0.0021, 0.00238)
	dur7	0.00232 (0.0021, 0.00255)	0.00223 (0.00209, 0.00238)
WLR[R(6)]	dur5	0.000544 (0.000349, 0.000718)	0.00169 (0.00158, 0.00178)
	dur6	0.00056 (0.000358, 0.000748)	0.00169 (0.00158, 0.0018)
	dur7	0.000564 (0.000369, 0.000763)	0.00169 (0.00157, 0.0018)

Table 5. Predicted values of end of trial Q-first moment, based upon a change of variables technique, and corresponding actual values from simulation

Statistic		Projected End of Trial Q 1st Moment (95% CI)	Actual End of Trial Q 1st Moment (95% CI)
WLR[R(3)]	dur5	0.00541 (0.00461, 0.00623)	0.00395 (0.00374, 0.00416)
	dur6	0.00533 (0.00459, 0.00615)	0.00395 (0.00373, 0.00415)
	dur7	0.00531 (0.0044, 0.00618)	0.00394 (0.00373, 0.00413)
WLR[R(4)]	dur5	0.00415 (0.00365, 0.00476)	0.00347 (0.00329, 0.00365)
	dur6	0.00407 (0.00356, 0.00459)	0.00348 (0.00328, 0.00365)
	dur7	0.00406 (0.00348, 0.00462)	0.00346 (0.00328, 0.00364)
WLR[R(5)]	dur5	0.00326 (0.00297, 0.00357)	0.00301 (0.00285, 0.00317)
	dur6	0.00322 (0.00294, 0.0035)	0.00302 (0.00285, 0.00318)
	dur7	0.00322 (0.00293, 0.00352)	0.00302 (0.00285, 0.00318)
WLR[R(6)]	dur5	0.00372 (0.00337, 0.00408)	0.00259 (0.00246, 0.00273)
	dur6	0.00368 (0.00331, 0.00402)	0.00259 (0.00245, 0.00273)
	dur7	0.00366 (0.0033, 0.00406)	0.00259 (0.00244, 0.00273)

The results in the first 8 lines of table 6 should be compared to this value. The procedure is intended to remove bias from the estimate in trials stopped early, and reduction in variability of the resulting estimate, if any would be an added benefit. The results of applying the bias adjustment procedure are surprising—there is little if any bias adjustment, conditionally upon the stopping time. The resulting estimator is still unbiased *unconditionally* upon the stopping time, but this is also the case for the unadjusted estimate. What we do gain upon applying the Rao-Blackwell technique is indeed a sizeable reduction in variability. One way to achieve this gain in precision would be to derive the design adjusted p-value and confidence interval for the unadjusted statistic on the standard normal scale and convert the upper and lower limits into the Rao-Blackwell-ized values.

5. THE NLST

The design of the NLST [11] 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 attainment of 1,200 lung cancer deaths, 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 constructed via the Lan-Demets procedure with a total probability of type II error set to 0.10. The drift at each interim analysis was derived under the $q \propto Q$ optimal weighting assumption and incorporated the design alternative $\beta^\star = \log(0.85)$. Initial estimates of $v(\tau)$ and $m(\tau)$ were posed in the design.

Table 6. Raw and bias adjusted estimates of $\beta^\star$ with simulated 95% CI according to the duration of trial. Each consecutive pair of lines shows the raw and then the bias adjusted estimate. Raw and bias adjusted estimates are shown corresponding to whether the predicted (P) end of trial variance or true (T) end of trial variance is used, whether the true shape is $Ramp[4]$ ($R[4]$) or Flat, and according to the monitoring statistic used, being $Ramp[3]$, $Ramp[4]$, $Ramp[5]$ or $Ramp[6]$ ($R[n]$)

variance, true shape, statistic	Duration 5	Duration 6	Duration 7
P,R[4],R[3]	-0.2 (-0.26, -0.16)	-0.15 (-0.2, -0.11)	-0.12 (-0.16, -0.085)
	-0.2 (-0.25, -0.17)	-0.13 (-0.15, -0.11)	-0.09 (-0.1, -0.08)
R[4]	-0.28 (-0.34, -0.23)	-0.19 (-0.26, -0.14)	-0.14 (-0.19, -0.1)
	-0.29 (-0.34, -0.25)	-0.18 (-0.2, -0.15)	-0.11 (-0.12, -0.1)
R[5]	-0.37 (-0.43, -0.32)	-0.24 (-0.32, -0.18)	-0.16 (-0.23, -0.11)
	-0.38 (-0.44, -0.35)	-0.22 (-0.26, -0.19)	-0.13 (-0.14, -0.12)
R[6]	-0.3 (-0.36, -0.27)	-0.19 (-0.26, -0.15)	-0.13 (-0.18, -0.088)
	-0.27 (-0.31, -0.22)	-0.17 (-0.32, -0.11)	-0.16 (-0.34, -0.055)
T,R[4],R[3]	-0.25 (-0.34, 0.0053)	-0.19 (-0.26, -0.051)	-0.14 (-0.2, -0.075)
	-0.27 (-0.32, -0.23)	-0.17 (-0.19, -0.16)	-0.12 (-0.12, -0.11)
R[4]	-0.31 (-0.41, 0.026)	-0.21 (-0.29, -0.051)	-0.15 (-0.22, -0.072)
	-0.34 (-0.4, -0.3)	-0.2 (-0.22, -0.18)	-0.13 (-0.13, -0.12)
R[5]	-0.37 (-0.47, -0.27)	-0.24 (-0.34, -0.19)	-0.15 (-0.23, -0.074)
	-0.41 (-0.47, -0.37)	-0.23 (-0.27, -0.2)	-0.13 (-0.14, -0.12)
R[6]	-0.41 (-0.5, -0.38)	-0.26 (-0.36, -0.2)	-0.16 (-0.24, -0.07)
	-0.44 (-0.5, -0.4)	-0.26 (-0.3, -0.22)	-0.14 (-0.15, -0.13)
P,Flat,R[3]	-0.2 (-0.26, -0.16)	-0.14 (-0.18, -0.1)	-0.11 (-0.15, -0.081)
	-0.2 (-0.25, -0.17)	-0.13 (-0.14, -0.11)	-0.087 (-0.097, -0.079)
R[4]	-0.29 (-0.37, -0.23)	-0.18 (-0.24, -0.14)	-0.13 (-0.18, -0.098)
	-0.3 (-0.36, -0.26)	-0.17 (-0.2, -0.15)	-0.11 (-0.12, -0.1)
R[5]	-0.39 (-0.47, -0.33)	-0.23 (-0.29, -0.18)	-0.15 (-0.21, -0.11)
	-0.4 (-0.47, -0.35)	-0.22 (-0.25, -0.19)	-0.13 (-0.14, -0.12)
R[6]	-0.32 (-0.38, -0.27)	-0.18 (-0.24, -0.14)	-0.12 (-0.16, -0.084)
	-0.27 (-0.32, -0.24)	-0.18 (-0.3, -0.11)	-0.16 (-0.33, -0.053)
T,Flat,R[3]	-0.24 (-0.36, 0.026)	-0.16 (-0.24, -0.022)	-0.12 (-0.18, -0.047)
	-0.28 (-0.33, -0.24)	-0.17 (-0.19, -0.15)	-0.12 (-0.12, -0.11)
R[4]	-0.29 (-0.43, 0.055)	-0.18 (-0.27, -0.013)	-0.12 (-0.2, -0.043)
	-0.34 (-0.42, -0.3)	-0.2 (-0.22, -0.18)	-0.12 (-0.13, -0.12)
R[5]	-0.37 (-0.5, 0.077)	-0.2 (-0.31, 0.0014)	-0.12 (-0.21, -0.04)
	-0.42 (-0.5, -0.37)	-0.23 (-0.26, -0.2)	-0.13 (-0.14, -0.12)
R[6]	-0.4 (-0.54, 0.091)	-0.22 (-0.33, 0.0029)	-0.13 (-0.21, -0.036)
	-0.47 (-0.54, -0.41)	-0.25 (-0.29, -0.22)	-0.14 (-0.15, -0.12)

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.

6. DISCUSSION

In prevention trials, or in general, trials in which we expect a delayed benefit, we have shown that use of a WLR statistic with weighting function of a Ramp form has desirable properties, even if the time at which the maximum weighting is attained is ill specified. We have proved that its distribution is asymptotically normal and shown how its values at early termination can be converted into a clinically meaningful parameter, the weighted average logged relative risk.

Our investigation into the power in our simulation study had a clear message. When a delayed benefit is expected, the WLR is the clear winner in terms of power as compared to the unweighted log-rank statistic, with gains in power observed constant even when the time to maximum value is misspecified, and losses in power when the shape function is constant on a par with losses observed when the true shape is ramped and the unweighted log-rank statistic is used. For the single point analysis this is certainly not news, but for an interim analysis schedule it is noteworthy, especially the gains in power despite mis-specification of the time to maximum weighting.

A possible drawback of the use of a WLR statistic in monitoring trials of the type investigated here is that the time to stop the trial in cases in which the intervention results in harm has a less favorable distribution than that of the unweighted log-rank statistic. When the size of harm is a 20% increased risk of event specific mortality, and the trial is monitored via a one sided efficacy boundary with a one sided futility boundary for monitoring against futility or harm, we saw that it took two analyses for termination probability in excess of 95% or more for the WLR statistic, whereas this probability was attained at the first analysis for the unweighted log-rank statistic. We argue that when there is little chance that the intervention, in this case screening, will result in harm in terms of the primary endpoint, here being cancer mortality, that this does not make the case against use of the WLR statistic for monitoring in prevention trials or similar trials in which we expect a delayed benefit. Certainly there is a possible downside to screening, which is one of the main reasons we have conducted large trials of cancer screening, but these downsides are realized in terms of increased morbidity or in the worst cases, other cause mortality, certainly not cancer specific mortality to any degree. Secondly, if there is substantial indication of harm, the DSMB will halt the trial regardless of whether the statistic has crossed a boundary or not. Therefore, the deficit in performance in terms of termination in case of harm should not cause too much concern in these cases.

As we saw in our simulation study, in some cases the predicted end of trial variance and Q-first moment were in error by an unacceptably large amount. We offer the following resolution. As shown in appendix 8, the end of trial functionals are predicted via a change of variables, whereby all quantities in each of the integrals are forced to depend only upon the pooled cumulative cancer mortality. In so doing, we assumed that the other cause mortality is proportional to the pooled cancer mortality, and that both the accrual and the increasing portion of the ramp weighting function are linear in the pooled cumulative cancer mortality. Violations to the first assumption are not critical at all since the effect is a misspecification of a survival function which is within less than 3% of unity. However, violations to the second and third assumption are more critical and are the source of the lack of acceptable margins of error in predicted values of these functionals observed in the simulated trials. The result of error in the end of trial variance will result in miscalculation of efficacy boundary points and could result in inflated type I error, while the result of error in end of trial first moment affects the conversion of WLR statistic to the estimate of $\beta^\star$, and the projected value of the end of trial drift which is required when stochastic curtailment is used in a futility analysis. These potential problems can be ameliorated by using a value of the end of trial variance obtained from the initial simulation study in the interim analysis plan, and using this value as the maximum information for the trial. This ensures that the end of trial variance is exact, thereby ensuring the exact type I error is maintained throughout monitoring of the trial. The end of trial first moment can also be stipulated in the interim analysis plan. This value can be obtained in the following way. In the simulation study, the ratio of the end of trial Q-first moment to the end of trial Q-second moment, was consistently within 6 percent of its mean value for each of the weighted statistics, regardless of the stopping time. Therefore, this ratio can

be obtained during the simulation study leading conducted during the design of the the interim analysis plan, thereby assuring that the estimate of $\beta^\star$ and end of trial drift, if required, are within 6% of their true values.

We have discussed several modifications to the standard methodology that are required when a sequentially monitored trial is monitored using a weighted log-rank statistic. We have seen that in the setting of a prevention trial when we expect that the effect of the intervention will be delayed by several years that use of a WLR statistic can be warranted. Use of a WLR statistic for monitoring a sequential trial comes with its own set of challenges. First, the value of the end of trial variance or total information should be specified in the interim analysis plan design. We are quite accustomed to trials being designed and powered on a requisite number of events (maximum information trials). Typically the initial design is in place and then an interim analysis plan is designed shortly thereafter, or perhaps even after the trial has begun but before any formal analyses are to take place. We remind the reader that the real target of the initial design and of the interim analysis plan is the end of trial variance or total information. If the investigators are comfortable running the trial until a pre-specified total information is attained then the best practice would be to conduct a interim analysis plan benchmarking study of the type outlined here and then predict the total information expected. Often, however, trials must be halted based upon calendar time because of funding or just the desire to have a concrete a-priori known end date. In these cases one should predict the total information expected at the pre-specified trial close-out date at the stage of interim analysis plan design.

A second challenge is the need for a shape assumption, that the chosen weighting function is proportional to the true logged relative risk function. This assumption becomes necessary in two possible ways. First if the design specifies a futility boundary, then the drift function must be predicted at each analysis. Since the interim analysis plan must allow an arbitrary timing of analyses, then either we must specify the drift function at all times in the initial design or we must make the above shape assumption in order to convert between a relative risk and values of the drift function during the run of the trial. The second way in which the shape assumption becomes necessary when it is desired that the reported relative risk and its design adjusted 95% confidence interval should correspond to the test of the null hypothesis in a one-to-one fashion. It is hard to imagine why the alternate case could possibly be argued at all. If there is to be such one to one correspondence between the sequentially tested null hypothesis and the reported relative risk and its 95% design adjusted confidence interval then we must have a mapping between values of the statistic on the monitoring (standard normal) scale and on the scale of the relative risk. While it is true in the general case, that the need to impose such a shape assumption is very severely limiting and unrealistic, we have tried to make a case for the argument that in the setting of prevention trials we have a-priori knowledge that the true shape of the logged relative risk is well approximated by a "ramp" function, and the time at which the maximum benefit is reached should be the time at which the preventive measure stops plus the mean lead time.

We have compared the performance of a monitoring plan based upon an unweighted log-rank statistic to that of a monitoring plan based upon a WLR. With the exception of stopping the trial early in the case that the intervention results in harm *in terms of the primary endpoint*, we have seen that the advantages of WLR statistic based monitoring greatly outweighs those of unweighted log-rank statistic based monitoring. We argue that if there is little chance that the intervention will result in harm in terms of the primary endpoint that this possible shortcoming of the WLR based monitoring can be overlooked. We also investigated the effect of mis-specifying weighting function (mean lead time) upon the performance of a WLR statistic based monitoring. In the simulations conducted here, the intervention was "offered" during the first two years of follow-up, and the maximal effect was realized in year 4, meaning that the true mean lead time was 2 years. The quantity most affected by mis-specification of the weight function is the early termination estimate of $\beta^\star$, and although we did not look at performance of the futiltiy monitoring, this would also affect estimates of the drift function and consequently the computed points on the futility boundary, but at least in our investigations, these effects did not seem to be quite so severe when the true shape was ramped, but when the true shape was flat, all ramp weighted statistics overestimate the true effect by roughly 20% or more. In future work it would be of interest to investigate the behavior of computed futility boundary points under mis-specified weights.

In conclusion, a WLR statistic based monitoring plan is recommended in the setting of a prevention trial, but careful attention must be paid to challenges presented by this type of monitoring, and one must be fairly sure that some delay in effect will occur, because if the true shape is flat there will be error in estimates of $\beta^\star$ and, if required, estimates of futility boundary. If there is a reasonable level of certainty that a delayed benefit will occur, then WLR statistic based monitoring can and should be used.

7. APPENDIX: LEMMAS 7.1 AND 7.2 AND PROOF OF THEOREM 2.1

In this appendix we present two lemmas and the proof of the main theorem, asymptotics of the scaled weighted log-rank statistic to a gaussian process with a drift. We begin with the compensated counting process. Based upon notation presented in section 2.1, the following difference is a compensated counting process martingale:

$$M_i(t) = dN_i(t) - \int_0^t I(T_i \geq \xi) \exp(Z_i \beta^\star q(\xi)) dH_0(\xi) \quad (26)$$

Put $M_{+,n}(t) = \sum_{i=1}^{n} M_i(t)$. The following two lemmas, which ensure that quantities used to rescale the score process converge in probability, are required in the proof of the main result, theorem 2.1. The first of these was also required in defining the cross moment presented in section 2.1. The first of these concerns the proportion at risk, the cumulative proportion of events and a related integral against increments in the cumulative proportion of events, and requires only boundedness of the cumulative baseline hazard and chosen weighting function.

Lemma 7.1. *Under condition 2.2, it follows that*

i.

$$R(\xi, 0) = \lim_{n\to\infty}{}_{a.s.}\, R_n(\xi, 0) \text{ exists,}$$

ii.

$$e(\xi, 0) = \lim_{n\to\infty}{}_{a.s.}\, E_n(\xi, 0) \text{ exists,}$$

iii.

$$G(t) = \lim_{n\to\infty}{}_{a.s.}\, N_n(t)/n \text{ exists.}$$

iv. and for any pair of functions ψ_1 and ψ_2 that are bounded on $[0, \tau]$, it follows that

$$\int_0^t \psi_1(\xi)\psi_2(\xi)e(\xi, 0)\left\{1 - e(\xi, 0)\right\} dG(\xi) \tag{27}$$
$$= \lim_{n\to\infty}{}_{a.s.} \int_0^t \psi_1(\xi)\psi_2(\xi)E_n(\xi, 0)\left\{1 - E_n(\xi, 0)\right\} \frac{dN_n(\xi)}{n}\,.$$

Proof. Parts (i) and (iii) follow by the strong law of large numbers as each of the corresponding expressions is an n-th partial sum, normalized by n, of an i.i.d. sequence of terms having common finite absolute first moment. Likewise, part (ii) follows by a similar argument, being the ratio of similarly described partial sums of i.i.d. sequences. Next, note that the second integral in part (iv) is defined path-wise as a sum. The existence and form of the almost sure limit follows path-wise almost-surely, from parts (ii) and (iii), due to the existence of the first written integral in the Rieman-Stiltjes sense. □

The second lemma concerns the risk weighted proportions at risk and an integral against increments in the compensated proportion of events. It requires absolute continuity between the trial arm specific cumulative hazard increment measures and boundedness of the instantaneous relative risk in addition to condition 2.2.

Lemma 7.2. *Let*

$$R_n(\xi, \theta) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq \xi)\exp(Z_i q(\xi)\theta)$$

and

$$E_n(\xi, \theta) = \frac{1}{nR_n(\xi, \theta)}\sum_{i=1}^{n} Z_i I(T_i \geq \xi)\exp(Z_i q(\xi)\theta)\,.$$

Under conditions 2.2, 2.1 and 2.3, it follows for any $\theta \in \mathbb{R}$ that

i.

$$R(\xi, \theta) = \lim_{n\to\infty}{}_{a.s.}\, R_n(\xi, \theta) \text{ exists,}$$

ii. and

$$e(\xi, \theta) = \lim_{n\to\infty}{}_{a.s.}\, E_n(\xi, \theta) \text{ exists,}$$

iii. and for any pair of functions ψ_1 and ψ_2 that are bounded on $[0, \tau]$,

$$\lim_{n\to\infty}{}_{pr} \int_0^t \psi_1(\xi)\psi_2(\xi)E_n(\xi, 0)\left\{1 - E_n(\xi, 0)\right\} \tag{28}$$
$$\times \left\{\frac{dN_n(\xi)}{n} - R_n(\xi, \beta^\star)dH_0(\xi)\right\} = 0\,.$$

Proof. Part (i) follows by the strong law of large numbers as it is an n-th partial sum, normalized by n, of an i.i.d. sequence of terms having common finite absolute first moment. Likewise, part (ii) follows by a similar argument, being the ratio of similarly described partial sums of i.i.d. sequences. To establish convergence to zero in part (iii), we factor $1/\sqrt{n}$ from the difference of differentials and note that the resulting integral is $1/\sqrt{n}$ times an integral that converges in distribution to a Gaussian martingale, by theorem (6.2.1) of Fleming and Harrington [4]. Therefore, convergence to zero in probability is established. □

Proof of Theorem 2.1. 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 $\sqrt{n}$. In the following, we suppress the dependence of $\beta^\star = b^\star/\sqrt{n}$, but keep in mind that $\beta^\star = O(1/\sqrt{n})$.

$$\begin{aligned}
U_n(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t Q(\xi)\left\{Z_i - E_n(\xi, 0)\right\} dM_i(\xi) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t Q(\xi)\left\{Z_i - E_n(\xi, 0)\right\} \\
&\quad \times I(T_i \geq \xi)\exp(Z_i q(\xi)\beta^\star)dH_0(\xi) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t Q(\xi)\left\{Z_i - E_n(\xi, 0)\right\} dM_i(\xi) \\
&\quad + \sqrt{n}\int_0^t Q(\xi)\left\{E_n(\xi, \beta^\star) - E_n(\xi, 0)\right\} \\
&\quad \times R_n(\xi, \beta^\star)dH_0(\xi)\,.
\end{aligned}$$

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

$$\begin{aligned}U_n(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t Q(\xi)\left\{Z_i - E_n(\xi,0)\right\}dM_i(\xi)\\ &+ \sqrt{n}\beta^\star\int_0^t Q(\xi)q(\xi)E_n(\xi,0)\\ &\times\left\{1-E_n(\xi,0)\right\}R_n(\xi,\beta^\star)dH_0(\xi)\\ &+ \sqrt{n}(\beta^\star)^2\int_0^t Q(\xi)q^2(\xi)E_n(\xi,\beta_1)(1-E_n(\xi,\beta_1))\\ &\times(1-2E_n(\xi,\beta_1))R_n(\xi,\beta^\star)dH_0(\xi)\,,\end{aligned}$$

where β_1 in the remainder term lies between $\beta^\star$ and zero. Before continuing, note that the remainder term in the Taylor expansion tends to zero almost surely as $n\to\infty$ since the integral is uniformly bounded almost surely and $\sqrt{n}(\beta^\star)^2$ tends to zero as $n\to\infty$. We ignore the remainder term in the following. Next, we normalize by $\sqrt{V_n(\tau)}$ and obtain:

$$\begin{aligned}X_n(t) &= \frac{1}{\sqrt{n\,V_n(\tau)}}\sum_{i=1}^{n}\int_0^t Q(\xi)\left\{Z_i - E_n(\xi,0)\right\}dM_i(\xi)\\ &+ \sqrt{\frac{n}{V_n(\tau)}}\beta^\star\int_0^t Q(\xi)q(\xi)E_n(\xi,0)\\ &\times\left\{1-E_n(\xi,0)\right\}R_n(\xi,\beta^\star)dH_0(\xi)\\ &= W_n(f_n(t;\tau)) + \frac{\langle Q|\mathbb{F}_n|q\rangle_t}{\sqrt{\langle Q|\mathbb{F}_n|Q\rangle_\tau}}\sqrt{n}\beta^\star\\ &+ \sqrt{\frac{n}{V_n(\tau)}}\beta^\star\int_0^t Q(\xi)q(\xi)E_n(\xi,0)\\ &\times\left\{1-E_n(\xi,0)\right\}\left\{\frac{dN_n(\xi)}{n} - R_n(\xi,\beta^\star)dH_0(\xi)\right\}.\end{aligned}$$

Consider the last line above. By the boundedness of integrands and intensities, it follows from theorem (6.2.1) of Fleming and Harrington [4], that the first term converges in distribution to a standard Brownian motion. Under the family of local alternatives, $\beta_n^\star = b^\star/\sqrt{n}$, it follows from lemma 7.1, that the second term is consistent to the drift given in expression 8 above. By lemma 7.2, the third term is consistent to zero. Therefore the result follows by Slutzky's theorem. □

8. APPENDIX: END OF TRIAL FUNCTIONALS

In this section we demonstrate how to project values of the variance $v(\tau) = \langle Q|\mathbb{F}|Q\rangle_\tau$, and the "first moment" $m(\tau) = \langle Q|\mathbb{F}|1\rangle_\tau$ at the scheduled end of study, τ. 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 $\theta = -dlog(S_{oth})/dH$ 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*

$$S_{lr}(\xi) = \frac{H(\tau)-H(\xi)}{H(\tau)-H(\tau-t_{er})}\wedge 1, \tag{29}$$

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.

$$Q(\xi) = \frac{1-\exp(-H(\xi)\wedge H(t_c))}{1-\exp(-H(t_c))}. \tag{30}$$

This is the $G^{0,1}$ member of the $G^{\rho,\gamma}$ family of weighting functions mentioned in the introduction, only here, it is stopped at its value at $t=t_c$. If the pooled cancer mortality grows at a nearly constant rate, then the above is nearly identical to $Q(t)=\frac{t}{t_c}\wedge 1$.

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. 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.

8.1 Variance at planned termination

$$\begin{aligned}v(\tau) = \langle Q|\mathbb{F}|Q\rangle_\tau &= \int_0^\tau Q^2(\xi)e(\xi,0)\left(1-e(\xi,0)\right)dG(\xi)\\ &= \int_0^\tau Q^2(\xi)e(\xi,0)\left(1-e(\xi,0)\right)S_{oth}(\xi)S_{lr}(\xi)S(\xi)dH(\xi)\,.\end{aligned} \tag{31}$$

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

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

These evaluate to:

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

respectively.

8.2 First moment at planned termination

(32)

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

Under assumptions 8.1, 8.2, 8.3, and 8.4, we again apply the change of variables, $\eta = H(\xi)$, to obtain

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

These evaluate to

$$J_1 = \frac{1}{4}\left\{\frac{1-\mathrm{e}^{-(\theta+1)(H(t_c)\wedge H(\tau-t_{er}))}}{\theta+1} - \frac{1-\mathrm{e}^{-(\theta+2)(H(t_c)\wedge H(\tau-t_{er}))}}{\theta+2}\right\},$$

$$J_2 = \frac{1}{4} I(t_c < \tau - t_{er})\left(1-\mathrm{e}^{-H(t_c)}\right) \times \frac{\mathrm{e}^{-(\theta+1)H(t_c)} - \mathrm{e}^{-(\theta+1)H(\tau-t_{er})}}{\theta+1},$$

$$J_3 = \frac{I(t_c > \tau - t_{er})}{4\left(H(\tau)-H(\tau-t_{er})\right)} \times \left\{\left(\frac{1}{1+\theta}\left[\left(H(\tau)-H(\tau-t_{er})\right)\mathrm{e}^{-(\theta+1)H(\tau-t_{er})} - \left(H(\tau)-H(t_c)\right)\mathrm{e}^{-(\theta+1)H(t_c)}\right] - \frac{1}{\theta+2}\left[\left(H(\tau)-H(\tau-t_{er})\right)\mathrm{e}^{-(\theta+2)H(\tau-t_{er})} - \left(H(\tau)-H(t_c)\right)\mathrm{e}^{-(\theta+2)H(t_c)}\right]\right) - \left(\frac{\mathrm{e}^{-(\theta+1)H(\tau-t_{er})} - \mathrm{e}^{-(\theta+1)H(t_c)}}{(\theta+1)^2} - \frac{\mathrm{e}^{-(\theta+2)H(\tau-t_{er})} - \mathrm{e}^{-(\theta+2)H(t_c)}}{(\theta+2)^2}\right)\right\}$$

$$J_4 = \frac{I(t_c < \tau)\left(1-\mathrm{e}^{-H(t_c)}\right)}{4\left(H(\tau)-H(\tau-t_{er})\right)} \times \left\{\frac{\left(H(\tau)-H(t_c\vee(\tau-t_{er}))\right)\mathrm{e}^{-(\theta+1)H(t_c\vee(\tau-t_{er}))}}{\theta+1} - \frac{\mathrm{e}^{-(\theta+1)H(t_c\vee(\tau-t_{er}))} - \mathrm{e}^{-(\theta+1)H(\tau)}}{(\theta+1)^2}\right\}$$

respectively.

8.3 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

$$G(\tau) = \int_0^\tau S_{oth}(\xi)S_{lr}(\xi)S(\xi)dH(\xi)\,, \tag{33}$$

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.

Received 3 December 2012

REFERENCES

[1] Anderson, G. L., Manson, J., Wallace, R., Lund, B., Hall, D., Davis, S., Shumaker, S., Wang, C.-Y., Stein, E. and Prentice, R. L. (2003). Implementation of the Women's Health Initiative Study Design. *Ann Epidemiol* **13** S5–S17.

[2] Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significnce tests on accumulating data. *Journal of the Royal Statistical Society, Series A* **132** 235–244. MR0250405

[3] Emerson, S. S. (1993). Computation of the uniform minimum variance unibiased estimator of a normal mean following a group sequential trialdiscrete sequential boundaries for clinical trials. *Computers and Biomedical Research* **26** 68–73.

[4] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York. MR1100924

[5] Gillen, D. L. and Emerson, S. S. Non-transitivity in a class of weighted logrank statistics under non-proportional hazards. *Statistics and Probability Letters* **77** 123–130.

[6] Izmirlian, G. (2014a). PwrGSD: Power in a Group Sequential Design R package version 1.172.

[7] Izmirlian, G. (2014b). Using PwrGSD–Package Vignette. http://cran.r-project.org/web/packages/PwrGSD/vignettes/.

[8] Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663. MR0725380

[9] Nan, B., Lin, X., Lisabeth, L. D. and Harlow, S. D. (2006). Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. *JASA* **101** 65–77. MR2252434

[10] National Lung Screening Trial Research Team, D. R. A., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., Gareen, I. F., Gatsonis, C., Marcus, P. M. and Sicks, J. D. (2011). Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med* **365** 395–409.

[11] The National Lung Screening Trial Research Team (2011). The National Lung Screening Trial: Overview and Study Design. *Radiology* **258** 243–253.

[12] R Core Team (2013). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[13] Wieand, S., Gail, M. H., James, B. R. and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76** 585–592. MR1040651

[14] Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-samplesurvival data. *Biometrika* **92** 1–17. MR2158606

Grant Izmirlian
Biometry Research Group
Division of Cancer Prevention
National National Cancer Institute
BG 9609 RM 5E130 MSC 9789
9609 Medical Center Dr
Bethesda, MD 20892-9789
USA
E-mail address: izmirlig@mail.nih.gov