Testing for a Treatment-by-Stratum Interaction in a Sequential Clinical Trial *

By Benjamin Yakir and W.J. Hall

The Hebrew University of Jerusalem and University of Rochester

We consider a two-arm sequential trial with two or more strata. The trial is monitored and terminated under the assumption of a common treatment effect (if any) in all strata. A secondary question at the end of the trial is: Does the treatment effect differ across strata—that is, is there a treatment × stratum interaction? We provide a test of the null hypothesis of no interaction—a test that recognizes the sequential stopping rule and allows for uneven accumulation of information in the various strata, a common case. In the case of two strata, and either a group-sequential design or a fully-sequential linear-boundary design, an optimal property of the test is derived. A computational algorithm is provided, and two examples summarized.

1. Introduction. We introduce and motivate this problem in terms of a survival-analysis based clinical trial with two strata, but similar considerations apply for other stratified sequential trials, as noted in Sections 5-7 below.

Consider a sequential clinical trial utilizing survival analysis and assuming a proportional hazards model, with two strata. Assume that log-rank statistics are computed separately for each stratum, either periodically or continuously. Denote the asymptotic versions of these statistics by $X_i(t)$, $i = 1, 2$, where t is a measure of total information accumulated at the time these statistics were computed.

The resulting two processes are time-transformed independent Brownian motions. In other words, X_1 and X_2 are independent, each has independent increments and

$$X_i(t) \sim N\big(\theta_i v_i(t), v_i(t)\big), \quad i = 1, 2.$$

The drift parameter θ_i—the log hazard ratio—is a measure of the treatment effect in stratum i, and $v_i(t)$ is (an estimate of) the associated partial information ($i = 1, 2$). Here, each $v_i(\cdot)$ is a positive and non-decreasing function, and $v_1(t) + v_2(t) = t$ for all t. For now, we assume that $v_1(t)$ and $v_2(t)$ are *not* proportional—information being accumulated quite 'unevenly' in one stratum relative to the other. The special case of proportionality, in which great simplification occurs, is summarized separately in Section 3. (For extension to the case in which one or the other $v_i(\cdot)$ may initially be zero, see the end of Section 2.)

The trial is monitored by inspecting the overall log-rank statistic $X(t) = X_1(t) + X_2(t)$, with a prescribed stopping boundary. The trial is concluded once this monitoring process hits a pre-specified boundary. The boundary is chosen at the design stage of the trial and can be, for example, an O'Brien-Fleming [11] type for a

*Research supported by grant R01 HL58751 from the National Heart, Lung and Blood Institute (USA)

group-sequential trial or have a triangular shape [14] for a fully-sequential trial. The specification of the boundary is such that the power criteria of the associated test is met at a given target value of a common treatment effect (assuming $\theta_1 = \theta_2$) while the significance level of the test is controlled.

A major assumption in the design, and in the analysis, of a stratified sequential trial is that there is no treatment × stratum interaction. (This assumption is not as critical in the proportionality case; see Section 3.) Formally, the assumption is that $\theta_1 = \theta_2 = \theta$, with θ the unknown common treatment effect. The monitoring process $X(\cdot)$ is then Brownian motion with drift θ, and the primary analysis is about this common treatment effect θ. But if this common-θ assumption does not hold, then the overall treatment effect is not well-defined. The power calculations at the design of the trial are not adequate in that power depends on both θ_1 and θ_2. Moreover, a positive effect in one stratum may be counteracted by a negative effect in the other stratum, possibly resulting in failure to reject the null hypothesis even though the new treatment may have important therapeutic value for specific sub-populations. (When $\theta_1 \neq \theta_2$, the expectation of $X(t)$ is no longer proportional to its variance, except as noted in Section 3.)

Clearly, at the end of the trial, one should investigate the validity of this assumption. A minimum requirement is that the data gathered in the trial are not contradictory to the null hypothesis of no interaction—in other words, that the null hypothesis is not rejected when testing for a treatment × stratum interaction.

Assume that the total information in the trial is T. In a sequential trial, T is a stopping time, hence random. In a nonsequential trial, T is a constant, or at least ancillary, and hence can be considered as constant. A natural candidate for a test statistic in this nonsequential setting is

$$(1) \qquad Z(t) = \frac{1}{\sigma(t)}\left(\frac{X_1(t)}{v_1(t)} - \frac{X_2(t)}{v_2(t)}\right) \qquad \text{with} \qquad \sigma^2(t) = \frac{1}{v_1(t)} + \frac{1}{v_2(t)},$$

evaluated at $t = T$. The null distribution of $Z(T)$, under the null hypothesis $\theta_1 = \theta_2$, is standard normal, but the mean is not zero under alternatives. An intuitive test procedure is to reject $\theta_1 = \theta_2$ in favor of $\theta_1 \neq \theta_2$ whenever $|Z(T)| \geq z_{\alpha/2}$ (the latter defining the upper $\alpha/2$ tail area of the standard normal distribution). Indeed, applying standard results for a two-parameter exponential family [7], it can be shown that this test is a *uniformly most powerful unbiased test of significance level* α, or UMPU(α) for short. (See also Section 3. An *invariance* discussion is in Section 7.)

In the sequential setting, however, the situation is more complex. The null distribution of $Z(T)$ is, in general, not normal [15]. Its mean, even under the null distribution, may not be zero, nor is its distribution typically symmetric. Nor is it even conditionally normal (given T) except under special circumstances. (Although $X(t) \perp Z(t)$ for each t, typically $\operatorname{cov}(X(s), Z(t)) \neq 0$ for $s < t$ since

$$\operatorname{cov}[X(s), Z(t)] = \frac{1}{\sigma(t)}\left(\frac{v_1(s \wedge t)}{v_1(t)} - \frac{v_2(s \wedge t)}{v_2(t)}\right).)$$

It follows that the test suitable for a nonsequential setting is not valid in the sequential setting. Nevertheless, it can be shown that a test based on $Z(T)$, adjusted to the sequential sampling design, may again possess important optimality properties, developed here in Section 2.

In Section 3, we deal with the special case of proportionality, and in Section 4, describe how the p-value for the optimal test can be computed. In Section 5, we consider the case of more than two strata, proposing a valid test but without optimality. In Section 6, we exemplify the method, and conclude with miscellaneous comments in Section 7. An appendix provides an example of output from software developed for this purpose.

This work was motivated by the Multicenter Automatic Defibrillator Implantation Trial (MADIT), sponsored by CPI/Guidant and administered at the University of Rochester [10]. This was a fully-sequential trial (weekly analyses), with a triangular design, comparing survival in patients randomized to receive an implanted defibrillator with survival in those on conventional medical therapy, in a specifically defined population of patients with coronary heart disease. After twenty months into this trial, a new version of the defibrillator became available—a transvenous one, not requiring the surgical operation of the original transthoracic model. The MADIT executive committee was willing to assume, for primary analyses, that the two versions of the defibrillator would be equally effective (discounting the small risk of operative mortality with the older version); but it seemed likely that patient recruitment would be affected when surgery was no longer required. Hence, a second stratum of patients was created, including all patients at each enrolling center once the center gained authorization to use the new device. Information continued to flow from both strata throughout much of the five-year duration of the trial, but at non-proportional rates. Monitoring and primary analyses were based on the overall (stratified) log-rank statistic, as described above.

Upon conclusion of the trial, a 'naive' analysis was carried out to evaluate whether the two versions of the defibrillator were equally effective, using the test statistic (1) but ignoring the sequential nature of the trial; no interaction was evident. Later, a valid p-value was computed, as described herein; see Section 6.

2. An optimal test with two strata.

We need two definitions.

(1) A fully-sequential stopping boundary for a Brownian motion is said to be *linear* if consisting of a linear upper boundary (positive intercept), a linear lower boundary (negative intercept), and a vertical boundary ($t = t_{max}$, say, for some $t_{max} \leq \infty$); if the slope b_U of the upper boundary exceeds the slope b_L of the lower boundary, then t_{max} is required to be finite (to assure $T < \infty$ *a.s.*). This class includes *sequential probability ratio tests* ($b_L = b_U$), possibly *truncated* ($t_{max} < \infty$), *triangular tests* ($b_L > b_U$), possibly *truncated*, and *restricted designs* ($b_L < b_U$ and $t_{max} < \infty$), popular in sequential clinical trial design (e.g., [14, 2]). For optimality claims, we confine attention to this class of fully-sequential designs, considered by Anderson [1] and Hall [5], but also consider group-sequential designs (including group-sequential modifications of triangular designs, as in [14]).

(2) Write $Z = Z(T)$ and $X = X(T)$, and consider an unbiased test $\phi_\alpha(Z, X, T)$ of significance level α for testing $\theta_1 = \theta_2$ against a two-sided alternative, defined on the support $\mathcal{S}$ of (Z, X, T). The test is said to be *truncation-adaptable unbiased (TAU)* if, for each truncation of T, to $T_c = T \wedge c$ ($c > 0$) say, there exists an extension of ϕ_α

to the support $\mathcal{S}_c$ of $\big(Z(T_c), X(T_c), T_c\big)$ which is also unbiased and of significance level α. (The test ϕ_α and its extension agree on $\mathcal{S} \cap \mathcal{S}_c$; the extension is only to (z, x)-values not attainable when $T = c$ but attainable when $T_c = c$.) The test is said to be *uniformly most powerful truncation-adaptable unbiased of significance level* α $\big(UMPTAU(\alpha)\big)$ if it is uniformly most powerful among such tests. This is the optimality sought for group-sequential designs. Its estimation counterpart was introduced in [9].

The interpretation of TAU tests is that, at any interim analysis time $t = c$ prior to stopping, a modification of the decision boundaries can be carried out, preserving its significance level and unbiasedness, *as if* the original design were truncated at $T = c$. A consequence is that the p-value associated with the test procedure upon stopping cannot depend on design characteristics beyond T—a desirable property since, in practice, such characteristics as the times of future analyses are not always precisely fixed in advance.

Now for the optimality theorem, for testing the null hypothesis $\theta_1 = \theta_2$ against $\theta_1 \neq \theta_2$:

THEOREM 1. *Consider the test $\psi_\alpha(Z, X, T)$, the indicator of the event $\{Z \leq z_1(X, T, \alpha)$ or $Z \geq z_2(X, T, \alpha)\}$. The conditional critical values $z_i(X, T, \alpha)$, $i = 1, 2$, are solutions to the equations*

$$\mathrm{P}_{\theta_1=\theta_2}(Z \notin (z_1, z_2)) \,|\, X, T) = \alpha, \tag{2}$$

$$\mathrm{E}_{\theta_1=\theta_2}(Z; Z \notin (z_1, z_2) \,|\, X, T) = \alpha \, \mathrm{E}_{\theta_1=\theta_2}(Z \,|\, X, T), \tag{3}$$

holding with probability one under the null hypothesis. Then the test ψ_α is an unbiased test of significance level α. Moreover, it is

(i) a UMPU(α) test when the trial has a fully-sequential linear design; (ii) a UMPTAU(α) test when the trial has a group-sequential design.

An optimal one-sided test is similarly available. Simply set $-z_1$ or $z_2 = \infty$ and omit constraint (3). But we focus on the more popular two-sided case.

To obtain a corresponding conditional p-value $p = p(z|x, t)$ at the observed values (z, x, t), write $\mu = \mu(x, t)$ for the conditional mean of Z in (3), and then (2) and (3) imply

$$\mathrm{E}_{\theta_1=\theta_2}[(Z - \mu)1_I(Z)|x, t] = 0 \tag{4}$$

with $I = (z_1, z_2)$. Now set z_1 or z_2 in (4) equal to z, according as $z \leq \mu$ or $\geq \mu$, solve (4) for the other z_i, and then $p = \mathrm{P}_{\theta_1=\theta_2}(Z \notin I|x, t)$. The needed conditional distribution is given in Section 4.

Proof. First we note that the statistic (Z, X, T) is sufficient in the two-parameter model, so we may confine attention to tests based on it.

Let $f_{\theta_1,\theta_2}(z, x, t)$ be the joint density (w.r.t. a suitable measure) of (Z, X, T) at (z, x, t). By the Likelihood Ratio Identity [12], and the independence of $Z(t)$ and

$X(t)$, it follows that

$$(5)\ f_{\theta_1,\theta_2}(z,x,t) = f_{0,0}(z,x,t) \times \exp\left\{z\delta(t) - \frac{1}{2}\delta(t)^2\right\} \times \exp\left\{x\theta(t) - \frac{t}{2}\theta(t)^2\right\}$$

where

$$\delta(t) = \frac{\theta_1 - \theta_2}{\sigma(t)} \qquad \text{and} \qquad \theta(t) = \theta_1\frac{v_1(t)}{t} + \theta_2\frac{v_2(t)}{t}$$

with $\sigma(t)$ in (1). Hence, on the one hand, (X,T) is seen to be sufficient for θ under the null hypothesis (under which $\delta(t) \equiv 0$ and $\theta(t) \equiv \theta$), and, on the other hand, the conditional distribution of Z, given the statistic $(X,T) = (x,t)$, is from a natural exponential family parametrized by $\delta(t)$. Specifically, dividing (5) by its integral w.r.t. z, we find

$$(6) \qquad f_{\theta_1,\theta_2}(z|x,t) = f_{0,0}(z|x,t)\, e^{z\delta(t)} / \, M(s|x,t)\Big|_{s=\delta(t)}$$

where $M(s|x,t)$ is the moment generating function of the null conditional distribution.

From standard theory on testing in exponential families [7], it follows that, given any (x,t), the UMPU conditional test of level α is characterized by (2) and (3). The given test, therefore, has unconditional significance level α and is unconditionally unbiased. Moreover, the test is more powerful than any other test for which the conditional significance level is identically α.

In the linear-boundary fully-sequential case, the statistic (X,T) is *boundedly complete* under the null hypothesis [8]. Hence an overall significance level α implies a conditional level α.

In the group-sequential case, the statistic (X,T) is not complete, but confining attention to *truncation-adaptable unbiased tests*, the same uniqueness implication may be seen to hold. The first step in an induction argument, very much as in [9], is to consider extending the UMPU(α) nonsequential test appropriate for $T \equiv t_1$ by partitioning the event $\{T = t_1\}$ into a stopping set $\mathcal{S}_1$ where still $T = t_1$ and a continuation set $\mathcal{C}_1$ where $T = t_2$. Let $\psi(z,x,t)$ be the test defined above for this $K = 2$ case, and let ψ' be an alternative unbiased test of significance level α that agrees with ψ on $\mathcal{S}_1$; write $h = \psi' - \psi$ and $h_0(t,x) = E_{\{\theta_1=\theta_2\}}[h|t,x]$. Then $h_0 = 0$ on $\mathcal{S}_1$ and, by the characterization of zero-mean statistics in [9], h_0 must be 0 a.e., implying that $E_0[\psi'|T,X] \equiv E_0[\psi|T,X] = \alpha$ a.s.—i.e., ψ' is also a test of *Neyman structure.* A similar argument provides an induction step, leading to the desired uniqueness conclusion. □

We required that both $v_i(t)$'s be everywhere positive. If instead $v_2(t) = 0$ for $t \leq t_0$, say, then we must impose the requirement that T exceed t_0. Otherwise, when $T \leq t_0$, there is no information available about θ_2 and hence about the difference $\theta_1 - \theta_2$. In a group-sequential design, we can insist on $v_i(t) > 0$ at each possible inspection time t. But in a linear design with a delayed start-up ($T \geq t_1$ for some $t_1 > t_0$), completeness of (X,T) is not generally known, and so optimality remains in doubt.

Incidentally, in the fully-sequential linear design case, the test ψ_α is seen to be TAU(α) and hence also UMPTAU(α).

3. The proportionality case. Now assume $v_1(t) = pt$ for some $0 < p < 1$ and all t, implying $v_2(t) = qt$ for $q = 1 - p$. Then $X(t)$ is generally a Brownian motion with drift $\theta = p\theta_1 + q\theta_2$, and the stopping boundaries are valid for testing hypotheses about θ whether or not there is a differential treatment effect.

Moreover, the process $Y(t) = \sqrt{t}\, Z(t)$ is also a Brownian motion, independent of the process $X(t)$, and with drift $\sqrt{pq}\,(\theta_1 - \theta_2)$; hence, the null distribution of $Z(T)$ is standard normal. It is no longer critical for the validity of the basic trial whether or not there is a differential treatment effect, especially if θ_1 and θ_2 have the same sign. And the test described for the nonsequential setting (after (1)) is seen to be UMPU(α). This is consistent with Theorem 1, which did not formally exclude this case.

This case provides some basis for ignoring the sequential stopping rule when the $v_i(t)$'s are not far from proportionality, as recommended by Whitehead [14]. However, see Section 7.

4. Calculating the conditional p-value. Here we outline an algorithm for computing, partly by Monte Carlo, the conditional p-value (defined after (4))—which depends on the conditional distribution of Z and on the observed z and (x,t). When $\theta_1 = \theta_2$, the conditional distribution is parameter-free, and so we choose $\theta_1 = \theta_2 = 0$.

In the continuous-time case, the null conditional distribution of Z given $(X,T) = (x,t)$ is a convolution of two independent random variables Y and W [15]; Y is mean-zero normal with variance

$$\tau^2(t) = 1 - \frac{1}{\sigma^2(t)} \int_0^t \left(\frac{\dot{v}_1(s)}{v_1(t)} - \frac{\dot{v}_2(s)}{v_2(t)}\right)^2 ds,$$

and W is a linear functional of the monitoring process $X(\cdot)$ given by

$$W(t) = \frac{1}{\sigma(t)} \int_0^t \left(\frac{\dot{v}_1(s)}{v_1(t)} - \frac{\dot{v}_2(s)}{v_2(t)}\right) dX(s).$$

Using the fact that $v_1(t) + v_2(t) = t$, and writing $v = v_1(t)$ and $\dot{v}(s) = \dot{v}_1(s)$, we find $\sigma^2(t) = t/[v(t-v)]$,

$$(7)\ \tau^2(t) = \sigma^2(t)\left\{v - \int_0^t \dot{v}(s)^2 ds\right\} \quad \text{and} \quad W(t) = \sigma(t)\left\{\int_0^t \dot{v}(s) dX(s) - \frac{vx}{t}\right\}.$$

In the discrete-time case, with inspections at $t_1 < t_2 < \dots$, we let $n(T)$ be the final inspection—that is, $t_{n(T)} = T$ (and set $t_0 = 0$). With (x,t) the observed value of (X,T), write $n = n(t)$. Then (7) becomes

$$(8)\ \tau^2(t) = \sigma^2(t)\left\{v - \sum_{j=1}^{n} \frac{\Delta_v(j)^2}{\Delta_t(j)}\right\} \text{and}\, W(t) = \sigma(t)\left\{\sum_{j=1}^{n} \frac{\Delta_v(j)\Delta_x(j)}{\Delta_t(j)} - \frac{vx}{t}\right\}$$

where

$$\Delta_t(j) = t_j - t_{j-1}, \quad \Delta_v(j) = v(t_j) - v(t_{j-1}), \quad \Delta_x(j) = X(t_j) - X(t_{j-1}).$$

(See [15].) Note that, in the proportionality case, $\tau^2 \equiv 1$ and $W \equiv 0$.

We now approximate the distribution of W by Monte Carlo, focusing first on the discrete-time case, with n inspections. To generate a single w-value, proceed as follows: Generate $n-1$ random variables, each conditionally normally distributed as implied by a discretized Brownian bridge from $(0,0)$ to (t,x), but truncated to the continuation region of the monitoring process. These values $(x_1, x_2, \ldots, x_{n-1})$ represent the positions of the monitoring process at times $t_1, t_2, \ldots, t_{n-1}$, under null hypothesis conditions. (This can be done in reverse order ($j = n-1, n-2, \ldots, 1$), with X_j given $(x_{j+1}, x_{j+2}, \ldots, x_n)$ being $N(r_{j+1}x_{j+1}, r_{j+1}\Delta_t(j+1))$ with $r_{j+1} = t_j/t_{j+1}$—'accepting' the value x_j only if it lies in the continuation region for $X(t_j)$, and continuing to generate such values until one is accepted.) Then w is given by (8).

Repeat this a large number M of times (perhaps $M = 100,000$), resulting in M w's, say $w_1, \ldots, w_M$, with average $\overline{w}$.

To obtain the distribution of Z, we need the convolution of this empirical distribution of W with $N(0, \tau^2)$. The resulting density for Z is

$$f(z) = \frac{1}{M}\sum_{j=1}^{M} \frac{1}{\tau}\phi\left(\frac{z - w_j}{\tau}\right)$$

with ϕ the standard normal density. Transform to $Z' = (Z - \overline{w})/\tau$, and note that (4) is equivalent to $\mathrm{E}[Z' 1_{I'}(Z') | x, t] = 0$ with $I' = (z_1', z_2')$, the interval of z'-values corresponding to $z \in I$. Carrying out the integration term-by-term, setting $w_j' = (w_j - \overline{w})/\tau$, the equation to be solved becomes

$$(9) \quad \frac{1}{M}\sum_{j=1}^{M} \left\{ w_j' \left[\Phi(z_2' - w_j') - \Phi(z_1' - w_j')\right] - \left[\phi(z_2' - w_j') - \phi(z_1' - w_j')\right]\right\} = 0$$

with Φ the standard normal distribution function. If $z' < 0$, set $z_1' = z'$, solve (9) for z_2', and then

$$(10) \qquad p = 1 - \frac{1}{M}\sum_{j=1}^{M} \left[\Phi(z_2' - w_j') - \Phi(z_1' - w_j')\right].$$

If $z' > 0$, set $z_2' = z'$ and solve (9) for z_1', with p again given by (10).

To solve (9), a Newton-Raphson iterative solution can be initiated with the assumption that the distribution of Z is symmetric, so that the two z_i's are equidistant from $\overline{w}$, or the two z_i''s equidistant from 0.

For the continuous-time case, first partition $[0, t]$ into a large number n of subintervals of length t/n each. Then proceed as in the discrete-time case.

The algorithm can be extended to allow for $v_2(t) = 0$ for $t \leq t_0$ by confining all computations to the time interval (t_0, T), valid when the probability of $T < t_o$ is negligible.

5. Several strata. If there are m (> 2) strata, with $X(t) = X_1(t) + \ldots + X_m(t)$ and $v_1(t) + \ldots + v_m(t) = t$, the natural test statistic in the nonsequential setting is the likelihood ratio test statistic

$$Z^2 = Z(T)^2 = \sum_{i=1}^{m} v_i(T) \left(\frac{X_i(T)}{v_i(T)} - \hat{\theta} \right)^2 \tag{11}$$

with $\hat{\theta} = X(T)/T$, the maximum likelihood estimate under the null hypothesis of no differential treatment effect ($\theta_1 = \ldots = \theta_m$). (This statistic is used in meta analyses; see, e.g., [4].) In the nonsequential setting, and in the proportionality case with all $v_i(t)$'s proportional, the null distribution of Z^2 is chisquare with $m - 1$ degrees of freedom. We propose using Z^2 also in the sequential setting without proportionality, rejecting the null hypothesis whenever Z^2 exceeds a critical value derived from the conditional distribution of Z^2 given (X, T). In contrast to the two-strata case, this test has no known optimality. (Indeed, in the two-strata case, formula (11) is the square of (1), so this test differs from the two-sided tests given earlier and is not unbiased.)

The conditional null joint distribution of $(X_1, X_2, \ldots, X_m)$, given $(X, T) = (x, t)$, can be represented as a convolution of a zero-mean multinormal vector $\underline{Y}$ with variance matrix $\Sigma = (\sigma_{ij})$, where

$$\sigma_{ij}(t) = v_i(t) 1_{i=j} - \int_0^t \dot{v}_i(s) \dot{v}_j(s) ds,$$

and an m-dimensional functional $\underline{W}$ of the monitoring process $X(\cdot)$ with coordinates $W_i(t) = \int_0^t \dot{v}_i(s) dX(s)$, $1 \leq i \leq m$ [15]. The conditional distribution of the resulting statistic Z^2 can be determined by a computational process similar to that used in the two-strata case, except that here it will be advisable to generate $\underline{Y}$, as well as $\underline{W}$, by Monte Carlo rather than carrying out the convolution mathematically. This will result in a Monte Carlo distribution of Z^2-values, against which the observed statistic can be compared.

6. Examples. The MADIT trial [10] was briefly described in Section 1. A total of 196 patients were enrolled with 98 in each of the two device-type strata. The trial terminated in favor of the defibrillator arm, with a p-value of 0.009 and an estimated hazard ratio of 0.46 (median unbiased)—determined by PEST software of Brunier and Whitehead [3].

This primary analysis was stratified, and based on the assumption that the two versions of the defibrillator were equivalent in their effects on mortality. Information accrual in the two strata was not at all proportional; $v_2(t)$, once positive, was very much like a convex quadratic function of t. Hence, a naive analysis (ignoring the stopping rule) is not justified.

The naive test for a differential treatment effect (interaction) yielded a p-value of 0.87 while the test described here yielded a p-value of 0.71. (The computation on a Sun Microsystems Ultra 5 workstation required 10 seconds; the output appears as an appendix.)

Our computer program also provides some summary statistics of the null conditional distribution of the test statistic Z. In the nonsequential case, and in the sequential case of Section 3, this distribution is standard normal; in contrast, in this example this distribution had mean -0.201 and standard deviation 0.968, but negligible skewness and excess kurtosis coefficients. Neither analysis showed any evidence of an interaction, but the non-zero mean could make a naive analysis misleading.

The distribution of Z under alternatives is not readily accessible. But if the conditional distribution were a shift of the null, the conditional power at $\delta = |\theta_1 - \theta_2|$ could be approximated by

$$pow(\delta) \approx \Phi\big(- z_\alpha - \delta/(0.968\sigma)\big) + \Phi\big(- z_\alpha + \delta/(0.968\sigma)\big)$$

with $\sigma = (v_1^{-1} + v_2^{-1})^{1/2} = 0.556$. (*Conditional power* refers to power at $\delta > 0$ of the test of $\delta = 0$, conditional on the stopping coordinates (x, t).) The conditional power to detect a four-fold difference in strata-specific hazard ratios ($\delta = \log 4$), with a 5% significance level, would be 30%. That this power is quite modest is not surprising: The MADIT trial was designed to be efficient for the primary hypothesis, and it stopped quite early with relatively little information accumulated. The conditional power would be increased in a longer running trial.

The test was also used in another recent fully-sequential trial with two strata [13]. In it, the p-value for a treatment effect in the primary analysis, assuming no differential effect in the two strata, was 0.0006, with a hazard ratio estimated to be 0.348. However, the two strata-specific pairs of Kaplan-Meier survival curves showed evidence of a treatment effect in only one of the two strata. The naive test for a differential treatment effect had a p value (two-sided) of 0.17; the corrected p-value was 0.09. (The null conditional mean was 0.222; information accumulation was *not* proportional.) Although the statistical significance was borderline, the researchers intend to recommend use of the treatment under study in only one of the two groups defining the two strata.

Use of the methods to test for a treatment × center interaction in a multicenter group-sequential trial will be reported elsewhere.

7. Final comments. (1) The methodology is not limited to sequential trials based on survival analysis. But such trials provide popular examples in which information might well be accumulated unevenly across strata. Indeed, if there was need for stratification—that is, the baseline survival curves differ across strata—then information, which is roughly proportional to counts of endpoints, will be accumulated in a non-proportional way across strata, even when recruitment occurs in a proportional way.

However, in any staggered-entry trial, it may well happen that recruitment in different strata does not occur in a proportional way, and this would lead to non-proportional accumulation of information in any kind of associated trial, whether based on measured responses on every subject or on censored-data survival analyses. For example, when stratifying by centers, it often occurs that some centers are phased in late and others are phased out early; or extra efforts my be introduced mid-way in a trial to stimulate recruitment in an under-recruited stratum (e.g., when strata are defined by gender or by race). We suspect that the non-proportional case may well be the usual one.

(2) In typical sequential trials, the power of the tests proposed here may be quite limited: a trial efficiently powered for a primary hypothesis may have little power for secondary questions. But if the secondary question is considered critical at the design stage, power for it could be enhanced. Estimation and confidence limits for the differential treatment effect $\delta = \theta_1 - \theta_2$ will be dealt with in a subsequent paper [6]. Tests for individual contrasts among multiple strata effects will be presented elsewhere.

(3) It should be understood that the word 'optimal' as used herein refers to optimality when observing a Brownian motion with drift. This is therefore only an asymptotic optimality when applied to a trial utilizing a log-rank-based survival analysis, or other asymptotically-normal statistics.

(4) From our experience with examples, it appears that the error in naive analysis is caused primarily by a non-zero null conditional mean of $Z(t)$; a second order effect is due to the conditional variance being slightly less than unity, but non-normality plays a negligible role. Such a shift in mean (whatever its sign) would reduce the p-value relative to the naive one. If a shift in mean generally occurs, the naive p is generally conservative. (Ability to deal with a Brownian bridge constrained by linear boundaries (7), or by a sequence of 'windows' (8), —even its mean path— could perhaps lead to analytical approximations; see [15].)

(5) The test for interaction can also be used by a monitoring committee at an interim analysis, conditioning on the current coordinates (x, t). (Unless close to a boundary, the naive test may be adequately accurate.) Some modification of the trial may be in order if sufficient evidence of a differential treatment effect is discovered.

(6) *Invariance* concepts are often used in nonsequential inference problems like these. Indeed, in the two strata case, the model for the two component Brownian motions is invariant under the group $\{g_c | c \in \mathcal{R}\}$ with g_c adding the function $cv_i(t)$ to $X_i(t)$ $(i = 1, 2)$, and the process defined by (1) is a *maximal invariant process* under this group. The induced group on the parameter space has $\delta = \theta_1 - \theta_2$ as a maximal invariant. If inference about δ were the only issue, a different design would have been used. A design aimed at inference about the common drift, which would have reasonable power properties for various other convex combinations of the two drift parameters if unequal, is not going to have an invariant stopping time T. As a consequence, the model for (Z, X, T) has no invariance structure; in particular, the stopped process $Z(T)$ does not have δ as its sole parameter (except in the proportionality case). And standard invariant testing concepts (see [7]) do not apply.

The same reasoning applies in the multiple-strata case. The process defined by (11) (with T replaced by t) is a maximal invariant process under the corresponding group of transformations, but the stopping time is not invariant.

(7) A Fortran program for computing p-values, suitable for both the two-strata and multiple-strata cases, may be obtained from the second author.

8. Acknowledgements. This work was developed jointly, and presented at the Conference by the first author. We thank Vladimir Dragalin and Aiyi Liu for technical assistance and helpful discussion, and we thank Arthur Moss for his eager

willingness in adopting a sequential design for the MADIT trial, which stimulated this research. We also acknowledge helpful suggestions from referees.

9. Appendix: Strata Program Output for the MADIT Example

```
Testing Equality of the First  2 Strata-Specific Treatment Effects

  number of strata =  2     number of data-inspections = 24

       stratum      x       v[t]     theta-hat
       -------   -------   -------  ---------
         1       7.8992    7.6733    1.0294
         2       5.2673    5.6033    0.9400

       overall   13.1665   13.2766    0.9917


   delta-hat =  0.089   sigma =  0.556   zobs =  0.161

       naive one-sided p-values = 0.564, 0.436
       naive two-sided p-value  =    0.872

 Statistics of the null conditional distribution

     of Z = delta-hat/sigma = Y (normal) + W:

          mean       sd        sk        kur
         ------    -----     -----     -----
     Y      0      0.808       0         0
     W   -0.201    0.533    -0.104    -0.015
     Z   -0.201    0.968    -0.017    -0.001


 Iterative solution for the other z-value:

         z-value   function  derivative
         -------   --------  ----------
     1  -0.56371   -0.00014   -0.1388
     2  -0.56451    0.00000   -0.1391

 The two z-values are:  -0.565,  0.161.


 The corrected one-sided conditional p-values are:  0.645, 0.355.
 The corrected two-sided conditional p-value is:      0.708.
```

```
# of Monte Carlo runs: 100,000      random numbers rejected: 4.0%
original seed =         212           start time: 17:34:12
final seed    =  2025285029             end time:   17:34:22
```

REFERENCES

[1] T.W. Anderson. A modification of the sequential probability ratio test to reduce the sample size. *Annals of Mathematical Statistics*, 31:165–197, 1960.

[2] P. Armitage. *Sequential Medical Trials.* Blackwell, Oxford, 2nd edition, 1975.

[3] H. Brunier and J. Whitehead. *PEST: Planning and Evaluation of Sequential Trials Operating Manual.* Medical and Pharmaceutical Statistics Research Unit, University of Reading, Reading, UK, 3.0 edition, 1993.

[4] W. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10:101–129, 1954.

[5] W.J. Hall. The distribution of Brownian motion on linear stopping boundaries. *Sequential Analysis*, 16:345–352, 1997. Addendum (1998) **17**, 123-124.

[6] W.J. Hall and B. Yakir. Inference about a secondary process following a sequential trial. *Institute of Mathematical Statistics Bulletin*, 29, 2000. abstract #266-205.

[7] E.L. Lehmann. *Testing Statistical Hypotheses.* Wiley, New York, 2nd edition, 1986.

[8] A. Liu and W.J. Hall. Minimum variance unbiased estimation of the drift of Brownian motion with linear stopping boundaries. *Sequential Analysis*, 17:91–107, 1998.

[9] A. Liu and W.J. Hall. Unbiased estimation following a group sequential test. *Biometrika*, 86:71–78, 1999.

[10] A.J. Moss, W.J. Hall, D.S. Cannom, J.P. Daubert, S.L. Higgins, H. Klein, J.H. Levine, S. Saksena, A.L. Waldo, D. Wilber, M.W. Brown, and M. Heo. Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. *New England Journal of Medicine*, 335:1933–1940, 1996.

[11] P.C. O'Brien and T.R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[12] D. Siegmund. *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York, 1985.

[13] S. A. Smith, D. J. Cobaugh, V. Dragalin, P. M. Wax, and R. A. Lawrence. An enhanced palatability charcoal in the management of pediatric poisoning victims. *Journal of Toxicology - Clinical Toxicology*, 67:611, 1999. abstract.

[14] J. Whitehead. *The Design and Analysis of Sequential Clinical Trials.* Wiley, New York, revised 2nd edition, 1997.

[15] B. Yakir. On the distribution of a concomitant statistic in a sequential trial. *Sequential Analysis*, 16:287–294, 1997.

Department of Statistics
The Hebrew University of Jerusalem
Jerusalem 91905, ISRAEL
msby@mscc.huji.ac.il

Department of Biostatistics
University of Rochester Medical Center
Rochester, New York 14642-8630, USA
hall@bst.rochester.edu