Published in final edited form as:

Pharm Stat. 2015 July ; 14(4): 273–283. doi:10.1002/pst.1683.

A novel test to compare two treatments based on endpoints involving both nonfatal and fatal events

Richard F. Potthoff^a and Susan Halabi^b

Richard F. Potthoff: potthoff@duke.edu

^aCancer Statistical Center, Duke University Medical Center, 2424 Erwin Road, Suite 802, Durham, NC 27705, U.S.A. phone: (919)681-1038

^bDepartment of Biostatistics and Bioinformatics, Duke University School of Medicine, 2424 Erwin Road, Suite 802, Durham, NC 27705, U.S.A

Abstract

In a clinical trial comparing two treatment groups, one commonly-used endpoint is time to death. Another is time until the first nonfatal event (if there is one) or until death (if not). Both endpoints have drawbacks. The wrong choice may adversely affect the value of the study, by impairing power if deaths are too few (with the first endpoint) or by lessening the role of mortality if not (with the second endpoint). We propose a compromise that provides a simple test based on the time to death if the patient has died, or time since randomization augmented by an increment otherwise. The test applies the ordinary two-sample Wilcoxon statistic to these values. The formula for the increment (the same for experimental and control patients) must be specified before the trial starts. In the simplest (and perhaps most useful) case, the increment assumes only two values, according to whether or not the (surviving) patient had a nonfatal event. More generally, the increment depends on the time of the first nonfatal event, if any, and the time since randomization. The test has correct Type I error even though it does not handle censoring in a customary way. For conditions where investigators would face no easy (advance) choice between the two older tests, simulation results favor the new test. An example using a renal-cancer trial is presented.

Keywords

Cancer progression; Composite endpoints; Log-rank test; Mortality; Two-sample Wilcoxon test

1. Introduction

The use of death as the endpoint in a clinical trial means that any nonfatal events, and the information they might provide, are disregarded. An alternative endpoint using nonfatal events, on the other hand, could be defined for cancer trials as the time of the first progression, recurrence, or relapse if one occurs or the time of death if not, but disregards any death following a nonfatal event. For instance, in a Phase III cancer trial where there may be interest in both overall survival and progression-free survival, the first endpoint fully disregards progressions and the second ignores deaths when prior progression is observed. Neither endpoint is invariably better with regard to a combination of power and clinical relevance.

The situation in cardiology trials is similar. A mortality endpoint by itself would be based on death before the end of the trial, or within a fixed time period. An alternative endpoint, a composite, could be based (e.g.) on rehospitalization, myocardial reinfarction, or death, whichever (if any) occurs first. Use of the mortality endpoint would lead to impaired power if deaths are fewer than expected. Use of the composite endpoint relies heavily on nonfatal events and could unduly downgrade the role of mortality.

In some trials, it may be obvious that either death or an endpoint related to nonfatal events is the better endpoint to use. Thus, mortality may be the appropriate endpoint if during the trial there will be enough deaths to provide good power and mortality is of leading importance. On the other hand, even if mortality would be an endpoint that is more meaningful clinically, the alternative endpoint may be the only practical one if it is likely that few patients will die before the trial ends.

There will be trials, though, where both types of endpoints are clinically relevant and could be measured but, at the design stage, one is uncertain about accrual rates, mortality rates, or nonfatal-event rates. Although mortality might ideally be the preferred endpoint, slower-than-expected rates for either patient accruals or deaths would mean that use of a mortality endpoint (necessarily decided upon in advance) would lead to lower power to detect differences. Conversely, use of a composite endpoint based on survival free of nonfatal events would be unfortunate if there are more deaths than expected, especially if mortality would be of greater clinical relevance.

It is at these uncertain situations that the present paper is directed. We propose a test that is a compromise between the two tests resulting from using a mortality endpoint alone and using a composite endpoint combining fatal and nonfatal events. We consider a randomized Phase III clinical trial where two treatment groups (experimental and control) are to be compared. Our null hypothesis is that the joint distribution of time from randomization until first (nonfatal or fatal) event and time from randomization until death is the same for the two groups.

Our test handles censoring in an unconventional way (by seemingly not recognizing that it exists) and so may appear, at first glance, to be violating basic statistical principles. Thus, to dispel misconceptions, we begin by considering the most unadorned form of the test (it also happens to be the form that, after refinement, may attract the most use). Let $I_0 > 0$ and $I_1 > 0$ be any two constants that are chosen in totally arbitrary fashion (*before* the trial starts). For both experimental and control patients, define the endpoint to be time to death if the patient has died; I_0 plus time from randomization (to data cutoff) if the patient has survived but had a nonfatal event; or I_1 plus time since randomization if the patient has survived without having a nonfatal event. Now apply the ordinary two-sample Wilcoxon test to compare the two groups using the endpoints as just defined. It should be clear that, regardless of the values of the increments I_0 and I_1 , the test will have the correct probability of rejection if the null hypothesis of no difference between the groups is true—that is, it will have correct Type I error.

Of course, power may be poor, depending on I_0 and I_1 . But Type I error is not impaired regardless of how I_0 and I_1 are (pre)determined. That remains true even if I_0 and I_1 are not necessarily constants, provided that their values, or the rules for getting them, are chosen before the trial begins (and are the same whether the patient is in the experimental or control group). Our test uses formulas to obtain I_0 and I_1 , rather than acquiring them as two arbitrary constants. In the most general form of our formulas, I_0 can vary with time since the first nonfatal event and I_1 with time since randomization, but in the simplest form I_0 and I_1 each turn out to be constant. To obtain I_0 and I_1 , the formulas calculate expected time until death using hypothetical (null) distributions with (pre-)specified form and parameters. Type I error is still maintained regardless of the degree of correctness of the hypothetical model and its parameters.

Thus the validity of our test does not depend on any distributional assumptions, nor are the increments I_0 and I_1 estimated from the trial data in any way. Although estimates of the hypothetical expected death times replace censored data wherever actual death times are censored, the means of obtaining those estimates are fixed before the trial starts. Our endpoints for censored patients are not to be considered as observations or as imputations. They merely provide a device for creating a valid test of the null hypothesis of no difference between the experimental and control groups: Disencumbrance from the complications of the censoring allows for proper application of the simple Wilcoxon test.

Section 2 explains the framework for our test. Formulas are given in Section 3. We ran power simulations comparing our test with two traditional competitors: the log-rank test with endpoint based only on the first nonfatal or fatal event (first nonfatal event if there is one, or death if not) and the log-rank test based only on death. Section 4 describes these simulations and Section 5 presents their results. (Part 2 of the Supplementary Materials covers further simulations.) Simulation outcomes for a simple and extreme special case that is unfavorable to our test, but still shows tolerable results, are given in Section 6. Section 7 presents an application of the method. Some general observations are in Section 8; Section 9 has concluding comments.

There is substantial work on problems pertaining to composite or multiple endpoints, including papers by Borm et al.,¹ Chi,² Lynch et al.,³ Quan et al.,⁴ Sandrini et al.,⁵ Song et al.,⁶ and Wong et al.⁷ However, none of those papers are concerned with the problem or approach covered in the present paper. Endpoints considered by previous authors often involved joint requirements on two or more components that make up a composite.

More generally, the areas of multi-state models, event-history analysis, and competing risks (e.g., Hougaard,⁸ Kalbfleisch and Prentice,⁹ Putter et al.¹⁰) bear some relation to the present paper. Our results, though, do not seem to overlap with anything previously found.

2. Framework for the test

Our test is based on a measure, Y , for each patient. That measure, which can be interpreted as the endpoint, is time of death if death has occurred, or expected time of death based on a pre-specified model if death is censored. We use variables Y , Q , T , and R with respective realizations y , q , t , and r . Define the following for each patient:

Q = Time from patient's randomization until data cutoff at the end of the trial (or until patient is lost to follow-up, if earlier)

T = Time from patient's randomization until death, censored if $t > q$

R = Time from patient's randomization until first event (which is a nonfatal event if $r < t$ or a fatal event if $r = t$), censored if $r > q$

Censoring of T is assumed to be independent of both T and R ; likewise for R . Let U_0 and U_1 denote the respective times when the first and last patients are randomized, and U_2 the time of data cutoff at the end of the trial. For a given patient, let U_E ($U_0 \leq U_E \leq U_1$) be the time of randomization, U_K the time of the first event, U_D the time of death, and U_L the earlier of U_2 and the time of any loss to follow-up. Then $q = U_L - U_E$; $t = U_D - U_E$, censored if $U_D > U_L$; and $r = U_K - U_E$, censored if $U_K > U_L$ (first event is nonfatal if $U_K < U_D$, fatal if $U_K = U_D$).

The measure Y for a patient is determined differently depending on whether (i) death has occurred ($t \leq q$); (ii) death is censored, but a nonfatal event is recorded ($r \leq q < t$); or (iii) death is censored, and no nonfatal event is recorded ($q < r \leq t$). Specifically, Y uses

- i. $y = t$ if $t \leq q$;
- ii. $y = E(t \mid q, r)$ if $r \leq q < t$;
- iii. $y = E(t \mid q)$ if $q < r \leq t$.

Thus (ii) and (iii) represent conditional expectations of time from randomization until death, conditional on the known values (q, r) and q , respectively. These conditional expectations are calculated from a hypothetical joint distribution of r and t , whose density is formulated as the marginal density of r times the conditional density of t given r (see Section 3).

Although $E(t \mid q, r)$ might more properly be written as $E(t \mid q, r, r \leq q < t)$ and $E(t \mid q)$ as $E(t \mid q, q < r \leq t)$, the abbreviated notations are used here for convenience.

Note that the measure y disregards any nonfatal events after the patient's first one. In fact, even the first nonfatal event (if any) is disregarded if death is observed.

The objective of our test is to test for differences between the two treatment groups, under the condition that no covariates are to be used. The null hypothesis is that the two treatment groups are the same with respect to the joint distribution of r and t . We use the two-sample Wilcoxon test to compare the measures y for the two groups, with the aim of achieving good power against alternatives where survival differs between the groups (though the null hypothesis does not refer to t alone).

We apply the ordinary Wilcoxon test, rather than the log-rank test, to the y 's because (mathematically) the y 's are all non-censored. Thus, the Wilcoxon test is applicable and one can benefit from its optimal properties (including its power, robustness, and accurate Type I error; see, e.g., Lehmann¹¹). The log-rank test is intended for censored data (though it is valid for non-censored data). Because in our simulations the log-rank test using the y 's performed poorly versus the Wilcoxon test using the y 's, our results for that particular use of the log-rank test are reported (briefly) in the Supplementary Materials (Part 2) rather than below.

3. Formulas for the test

As indicated in Section 2, the conditional expectations $E(t | q, r)$ for a patient who falls under case (ii) ($r < q < t$) and $E(t | q)$ for one who falls under case (iii) ($q < r < t$) are evaluated via a hypothetical joint distribution $f(r, t)$ that is expressed as $f(r, t) = f_1(r) f_2(t|r)$. It is the same for both treatment groups. Although the power of our test is affected by the choice of f_1 and f_2 , Type I error will be correct regardless of that choice and regardless of how closely it conforms to the true distributions that f_1 and f_2 are intended to represent. We base our choice on Weibull distributions, though, because they are realistic and versatile as well as mathematically tractable.

Thus we set

$$f_1(r) = a^b b r^{b-1} e^{-(ar)^b}, \quad 0 < r < \infty, \quad (1)$$

where the parameters a and b (as well as the choice of the Weibull distribution itself) have to be specified before the trial starts. For $f_2(t|r)$ the formulation is more complicated. Let p ($0 < p < 1$) denote the probability that the first event is nonfatal. The conditional distribution f_2 uses the continuous density

$$f_3(t|r) = g^h h (t-r)^{h-1} e^{-[g(t-r)]^h}, \quad r < t < \infty \quad (2)$$

with probability p , together with the discrete density that has the rest of the mass concentrated at the single point $t = r$ with probability $(1 - p)$. The values of p , g , and h all have to be specified (again in advance of the trial).

From (2) we find that the conditional expectation for case (ii) is

$$\begin{aligned} E(t|q, r) &= \frac{\int_q^\infty t f_3(t|r) dt}{\int_q^\infty f_3(t|r) dt} = \frac{\int_q^\infty t g^h h (t-r)^{h-1} e^{-[g(t-r)]^h} dt}{\int_q^\infty g^h h (t-r)^{h-1} e^{-[g(t-r)]^h} dt} \\ &= \frac{\int_{[g(q-r)]^h}^\infty \left(r + \frac{1}{g} u^{\frac{1}{h}}\right) e^{-u} du}{\int_{[g(q-r)]^h}^\infty e^{-u} du} = r + \frac{\int_{[g(q-r)]^h}^\infty u^{\frac{1}{h}} e^{-u} du}{g e^{-[g(q-r)]^h}}. \end{aligned} \quad (3)$$

In the simple case $h = 1$, (3) is just equal to $[q + (1/g)]$, so that $I_0 = 1/g$ and $E(t | q, r)$ is the time in the study plus $1/g$. Even if $h \neq 1$, (3) is still easily calculated, using an incomplete gamma function. [If $h \neq 1$, the increment I_0 is the value in (3) minus q .]

The conditional expectation for case (iii), $E(t | q)$, is obtained by using both (1) and (2). To start, note that

$$\begin{aligned} E(t|r) &= (1-p)r + p \int_r^\infty t f_3(t|r) dt \\ &= (1-p)r + p \int_0^\infty \left(r + \frac{1}{g} u^{\frac{1}{h}}\right) e^{-u} du = r + \frac{p}{g} \Gamma\left(1 + \frac{1}{h}\right). \end{aligned} \quad (4)$$

From (4) one obtains

$$E(t|q) = \frac{\int_q^\infty E(t|r) f_1(r) dr}{\int_q^\infty f_1(r) dr} = \frac{\int_q^\infty \left[r + \frac{p}{g} \Gamma\left(1 + \frac{1}{h}\right) \right] a^b b r^{b-1} e^{-(ar)^b} dr}{\int_q^\infty a^b b r^{b-1} e^{-(ar)^b} dr}. \quad (5)$$

Thus

$$E(t|q) = \frac{\int_{(aq)^b}^\infty v^{\frac{1}{b}} e^{-v} dv}{a e^{-(aq)^b}} + \frac{p}{g} \Gamma\left(1 + \frac{1}{h}\right), \quad (6)$$

an expression that reduces to $[q + (1/a) + (p/g) \Gamma(1 + \frac{1}{h})]$ if $b = 1$, or just to $[q + (1/a) + (p/g)]$ if $b = 1$ and $h = 1$ [so that $I_1 = (1/a) + (p/g)$ in this last case].

If desired, one can allow either p or g to vary with r . See Part 1 of the Supplementary Materials for details.

Our test can be summarized as follows. The value of y for each patient in each treatment group is simply the time until death (t) if death has occurred [case (i), $t \leq q$]. It is the hypothetical expected time until death, $E(t | q, r)$ from formula (3), if the patient is alive (at cutoff time) but has had a nonfatal event [case (ii), $r \leq q < t$]. It is the hypothetical expected death time given by $E(t | q)$ from formula (6) if the patient is alive and has not had a nonfatal event [case (iii), $q < r \leq t$]. Once the y 's are found for all the patients, the Wilcoxon two-sample rank-sum test is applied to these y -values to determine if there is a statistically significant difference between the two treatment groups.

4. Description of the simulations

The two main competitors of our test are the log-rank test based on the time to the first (nonfatal or fatal) event as the endpoint, hereafter denoted as the LRNF test; and the log-rank test based on time of death as the endpoint, hereafter denoted as LRF. We ran simulations to compare our test (hereafter W, for Wilcoxon) with LRNF and LRF under different conditions. Under LRNF, an observation is censored if neither a nonfatal event nor death has occurred by the cutoff time. Under LRF, an observation is censored if death has not occurred.

Certain elements, chosen to reflect reasonably common design specifications, are constant throughout our simulations. All calculations used SAS (including SAS/IML). There are two groups of patients, designated as treatment and control, with 300 patients in each. Enrollment lasts for 5 time units. The data cutoff is at 7 time units. For the 300 patients in each group, q follows a uniform (rectangular) distribution from 2 to 7. Such a specification would yield a power of 0.9 to detect a hazard ratio of 0.7 if the median time to event in the control group is 3 time units, assuming exponential survival times (log-rank test, two-sided $\alpha = 0.05$).

Because of space considerations, we report some of our simulations here and others in Part 2 of the Supplementary Materials. The former vary the true but not the hypothetical distributions of r and t (where “true” means true for the simulation). By contrast, the latter

simulations (except for one for the null case) vary both. Here we use a single set of (Weibull-based) hypothetical distributions defined by $f_1(r)$, $f_2(t|r)$, and $f_3(t|r)$ [see (1) and (2)] in terms of the five parameters a , g , b , h , and p . Only g and h affect (3), but the calculation of (6) uses all five parameters. In the descriptions that follow, though, we use the median time from randomization until a first (fatal or nonfatal) event, $A = (\log 2)^{1/b}/a$, in place of a ; and the median time from a nonfatal event until death, $G = (\log 2)^{1/h}/g$, rather than g .

Our set of hypothetical distributions used $A = 3$, $G = 3$, $b = 1$, $h = 1$, and $p = 0.75$ (same for both treatment and control). The choices for the hypothetical medians A and G would generally represent their anticipated values (for the control group, that is, or, equivalently, for the treatment group if the treatment were to have no effect, as would be the case under the null hypothesis). Likewise, the values chosen for b , h , and p would be in line with anticipations. The five values were chosen with an eye toward creating a typical situation where it could be hard to decide between LRNF and LRF.

The simulations have to specify a true joint distribution of r and t for the treatment and control groups separately. We let these true joint distributions take the same (Weibull-based) form as for the hypothetical distribution [see the two preceding paragraphs, plus (1) and (2)] but with parameters to be denoted by A^* , G^* , b^* , h^* , p^* and A^{**} , G^{**} , b^{**} , h^{**} , p^{**} for the control and treatment groups, respectively (in place of A , G , b , h , p).

We simulated three values for each of the five control-group parameters, as follows:

$$A^* = 0.8A, A, 1.2A = 2.4, 3.0, 3.6$$

$$G^* = 0.8G, G, 1.2G = 2.4, 3.0, 3.6$$

$$b^* = 7/8, 1, 8/7$$

$$h^* = 7/8, 1, 8/7$$

$$p^* = 0.65, 0.75, 0.85$$

The five treatment-group parameters had a fixed relation to the control-group parameters: $A^{**} = 1.32A^*$, $G^{**} = 1.32G^*$, $b^{**} = b^*$, $h^{**} = h^*$, and $p^{**} = p^* + 0.012$. Although two of these parameters are the same for control and treatment, the other three are not.

The true control medians, A^* and G^* , vary around the hypothetical median (A or G) so that the simulation can assess the effect of the true median turning out to be lower than, equal to, or higher than the hypothetical (or anticipated) median. The Weibull true shape parameters likewise take values below, equal to, and above the hypothetical values (of $b = 1$ and $h = 1$). The true control value for p^* (probability that the first event is nonfatal) is below, equal to, and above the hypothetical or anticipated value of 0.75.

The design for the simulations thus uses $3^5 = 243$ conditions for the true distributions. Each condition was run with 1000 replications, thereby yielding $243 \times 1000 = 243,000$ simulation outcomes in total (based on $243,000 \times 2 \times 300 = 145,800,000$ patients). Each outcome has test results for W, LRNF, and LRF.

The simulation parameter values were chosen to try to get LRNF and LRF to have similar power to each other overall (which is what happened). Such a situation corresponds to conditions that are generally most suitable for the use of W (conditions where both LRNF and LRF may be sub-optimal).

For each of W, LRNF, and LRF, the test statistic that we record for any of the 243,000 simulation outcomes is a value z that, under the null hypothesis, would follow approximately a normal distribution with mean 0 and variance 1. We calculate z as the square root of a χ^2 -value. For W, that χ^2 -value is the Kruskal-Wallis χ^2 for the Wilcoxon test from PROC NPAR1WAY of SAS. For LRNF and LRF, the χ^2 -value whose square root is set equal to z is from PROC LIFETEST of SAS. Neither the χ^2 for W nor the one for LRNF or LRF involves a continuity correction. A minus sign is attached to any z if (as does rarely happen) the test based on that z shows the control group with a better result than the treatment group.

We use z as a gauge of power. Because it is a monotonic function of the one-tailed significance level at which rejection of the null hypothesis would occur, it is comparable across the three tests. In addition, z is a continuous measure and is thus more sensitive than (e.g.) a binary measure that is based on whether or not the test rejects at an arbitrary level such as 0.025.

The results presented in the following section all use z rather than χ^2 to compare power of the three tests. Of course, the two give the same answer (with recognition of the minus signs) for simple ordinal comparisons. But for other results z should be a better gauge than χ^2 , because one can expect z to be less skewed and not as sensitive to outliers.

5. Results from the simulations

The simulation results fall into three main categories and apply within the limits of the conditions under which they were obtained. First, W performs well overall (with respect to power) and is competitive in relation to both LRNF and LRF. Second, when any of the five true parameters A^* , G^* , b^* , h^* , p^* differs from the corresponding hypothetical or anticipated parameter (A , G , b , h , p), the resulting misjudgment may not be damaging; specifically, either a negative true-minus-hypothetical discrepancy hurts W (in relation to LRNF or LRF) and a positive one helps, or vice versa. Third, use of W is apparently a relatively safe strategy in the sense that W does not often do worse than both LRNF and LRF. Details follow.

Of the total of 72,900,000 control-group patients (spread across different conditions), 36.9% experienced no event (either nonfatal or fatal) and 62.8% did not die. The corresponding respective censoring values for the treatment-group patients are 46.4% and 72.3%.

Of the 243,000 values of the difference ($z_W - z_{LRNF}$), 67.0% are positive. Their average is +0.231. Corresponding respective figures for the difference ($z_W - z_{LRF}$) are 67.5% and +0.232. [For the difference ($z_{LRNF} - z_{LRF}$), 49.8% of the values are positive.]

W also performed well with respect to the percentage of the 243,000 outcomes in which each of the three tests rejected the null hypothesis at the 0.025 and 0.005 levels (based on

one-tailed probabilities, with rejection only in the upper tail: $z > 1.960$ and $z > 2.576$). The respective rejection frequencies at the 0.025 and 0.005 levels are 80.2% and 60.0% for z_W , 73.1% and 51.2% for z_{LRNF} , and 73.6% and 51.2% for z_{LRF} .

At a much more detailed level, Tables 1 and 2 exhibit, for each of the 243 conditions, the percentage of the 1000 outcomes for which, respectively, $z_W > z_{LRNF}$ and $z_W > z_{LRF}$. These percentages exceed 50.0% for all but two of the 243 conditions in Table 1 and all but eight in Table 2. They exceed 53.1% for all but eight conditions in Table 1 and all but 17 in Table 2 (where 53.1% is the smallest percentage of successes that is exceeded less than 2.5% of the time in 1000 Bernoulli trials with success probability $\frac{1}{2}$). Thus W shows better power than both LRNF and LRF over a broad set of conditions among those that we simulated.

As additional information complementary to Tables 1 and 2, all but two of the 243 conditions have a positive average value (across the 1000 replications) for $(z_W - z_{LRNF})$. All but eight have a positive average for $(z_W - z_{LRF})$. For no condition are both averages negative.

We ran analyses of variance using the $(z_W - z_{LRNF})$ and $(z_W - z_{LRF})$ differences as dependent variables and first allowing for all 31 main effects and interactions among the five (true) parameters (degrees of freedom = 242 for the model, 242,757 for error). Of the 32 ($= 2 \times 16$) interactions of more than two factors, only four, all of minor magnitude, were significant at the 0.05 level. We then ran both analyses using only main effects and two-factor interactions (degrees of freedom = 50 for the model, 242,949 for error). Those results showed statistical significance for some two-factor interactions but little practical significance, whereas all main effects had marked significance. It is therefore reasonable to examine only the main effects, which are shown in Table 3 for each of the two differences at each of the three levels of each of the five parameters. Each number is the average of 81,000 values of $(z_W - z_{LRNF})$ or $(z_W - z_{LRF})$ for the indicated parameter level.

Of the 30 values in Table 3, none are negative, thus indicating some consistency in the favorable performance of W . An important question is whether W is harmed (does not perform as well versus LRNF or LRF) when the true parameters A^*, G^*, b^*, h^*, p^* turn out to differ from their anticipated (hypothetical) counterparts A, G, b, h, p that are specified in advance. Table 3 provides evidence concerning the effects of such misjudgment. The simulation has each of A^*, G^*, b^*, h^*, p^* at three levels ($<$, $=$, and $>$) in relation to A, G, b, h, p . If an $=$ result were better for W , in relation to LRNF or LRF, than *both* the $<$ and the $>$ results, that would not bode well for W (because W would then be performing worse under true-versus-hypothetical discrepancies in both directions). But such a result does not occur. Thus, Table 3 indicates that a discrepancy hurts in one direction but helps in the other; it does not hurt in both directions. In this sense, the average success of W in relation to LRNF or LRF does not depend (at least within the limits of the simulation) on accurately selecting the five hypothetical parameters for the control group so that they are close to their true values.

More specifically, Table 3 shows a triple of values (corresponding to $<$, $=$, and $>$) related to each of the five hypothetical control parameters, for each of LRNF and LRF in relation to

W. All of these 10 triples are monotonic. Main-effects analyses of variance (degrees of freedom = 10 for the model, 242 989 for error), run separately for each of $(z_W - z_{LRNF})$ and $(z_W - z_{LRF})$, find that linear effects are significant at the 0.0001 level in all 10 cases.

For each of W, LRNF, and LRF, Table 4 shows the percentage of the 243,000 outcomes in which the test places best (based on z), second best, and worst among the three tests. Only rarely—in 4.1% of the outcomes—is W worst. W is often second best, at 57.2%. The results of Table 4 suggest that W will be a good compromise test choice if the investigator is wary about both LRNF and LRF: W is not often worse than both, and thus provides a relatively safe choice.

6. Use of the test for an extreme special case

Although our test is aimed at a situation where both nonfatal and fatal events are of potential interest for defining endpoints, one could ask whether a form of the test might be used in a setting where there is only one type of (declared) endpoint, such as death. Such a setting corresponds to a degenerate case in the model of Sections 2 and 3, such that the probability (p) that the first event is nonfatal is taken to be 0 (all events are deaths, and nonfatal events are either nonexistent or disregarded). For this relatively simple setting it is possible to formulate a degenerate form of our test, as will be shown shortly. But one would be disinclined to use it in practice, because the setting is the standard one where the log-rank test is the usual choice and has optimal properties. Nonetheless, the setting may be close to a worst case for judging our test, thus suggesting that simulations for this extreme case can shed light on how much one might be hurt by choosing to use our test under conditions that turn out to be unexpectedly unfavorable.

The framework for our test for the special case can be set forth as follows. Let

Q = Time (as before) from patient's randomization until trial cuts off

R = Time from patient's randomization until event (death), censored if $r > q$

The measure Y for a patient is:

- i. $y = r$ if $r < q$ (event not censored)
- ii. $y = E(r | q)$ if $q < r$ (event censored)

There is, of course, no p , $f_2(t|r)$, $f_3(t|r)$, or $f(r, t)$, but only a (hypothetical) density $f_1(r)$ whose form is the same as (1) and whose parameters a and b again have to be (pre-)specified. The formula for (ii) is just

$$E(r|q) = \frac{\int_q^\infty r f_1(r) dr}{\int_q^\infty f_1(r) dr} = \frac{\int_q^\infty a^b b r^b e^{-(ar)^b} dr}{\int_q^\infty a^b b r^{b-1} e^{-(ar)^b} dr} = \frac{\int_{(aq)^b}^\infty v^{\frac{1}{b}} e^{-v} dv}{a e^{-(aq)^b}}, \quad (7)$$

which becomes $[q + (1/a)]$ if $b = 1$.

Our simulations for the special case pit our test, W, based on y with (7) applied for (ii), against the log-rank test, now to be denoted just by LR since LRNF and LRF are no longer

distinct tests. As before, the treatment and control groups have 300 patients each, enrollment ends after 5 time units, the cutoff of data comes after 7 time units, and q is drawn from a uniform distribution from 2 to 7. Throughout the simulations, the hypothetical distribution used for r is Weibull, as in (1), with median $A = (\log 2)^{1/b/a} = 3$ and shape parameter $b = 1$. The only true parameters that vary are A^* and $b^* = b^{**}$, which have the same meaning and take on the same values as in Section 4: $A^* = 0.8A$, $A, 1.2A = 2.4, 3.0, 3.6$; and $b^* = b^{**} = 7/8, 1, 8/7$. They generate $3^2 = 9$ conditions, each of which was run with 1000 replications to provide $9 \times 1000 = 9000$ simulation outcomes in total. The parameter A^{**} is set to $1.32A^*$, as before. The test statistic z , defined as in Section 4, was recorded for each of W and LR for each outcome.

Simulation results were as follows. There is censoring (i.e., no event occurred) for 36.9% of the control patients and 46.4% of the treatment patients. The average $(z_W - z_{LR})$ difference across all 9000 outcomes is -0.069 ; z_{LR} exceeds z_W in 62.9% of those outcomes and $z_W > z_{LR}$ in the remaining 37.1%. Upper-tail rejection occurs at the 0.025 level in 73.1% of the 9000 outcomes for LR and 71.2% for W, and at the 0.005 level in 51.2% for LR and 48.5% for W.

Respective main-effect estimates [averages of $(z_W - z_{LR})$] are $-0.092, -0.058, -0.059$ for $A^* = 2.4, 3.0, 3.6$; and $-0.059, -0.066, -0.083$ for $b^* = b^{**} = 7/8, 1, 8/7$. Among the 9 conditions of $(A^*, b^* = b^{**})$ combinations, the average $(z_W - z_{LR})$ difference ranges from -0.047 to -0.086 except for $(2.4, 8/7)$, whose value is -0.120 .

Comparison of these differences versus those in Table 3, and especially of the -0.069 overall difference in this section versus the $+0.231$ and $+0.232$ overall differences found in Section 5, indicates that even our extreme special case shows only a limited weakness of W in relation to the benefits of W under the conditions at which W is aimed. Thus the differences just presented, which could be realized if there were an unlucky guess to use W, are small enough to be of little concern in view of the size of the benefits shown in Section 5 for conditions where W could be expected to be used.

7. Example

We used data from Cancer and Leukemia Group B study 90206, a Phase III clinical trial of metastatic renal-cell carcinoma (see Rini et al.^{12, 13} for full details), to provide an example to compare our W test with LRNF and LRF. The trial was designed for 588 deaths so that the log-rank statistic would have 86% power to detect a hazard ratio of 0.76 assuming a two-sided significance level of 0.05. The study randomized 732 patients, 369 to a treatment group and 363 to a control group. The first and last patients were randomized on October 30, 2003 and July 1, 2005, respectively. The data cutoff date for the final analysis was March 24, 2009. Stratification factors that were applied for the trial itself were disregarded for purposes of this example. Although overall survival was specified as the primary endpoint for the trial, investigator-assessed progression-free survival was a secondary endpoint.

From trial documentation and design we judged that, before the start of the trial, $(143, 289, 1, 1, 0.875)$ were values that reasonably could have been hypothesized or anticipated for the control-group parameters (A, G, b, h, p) . (The median values A and G are expressed in days.)

For these same five parameters, values as estimated from the data are (196, 402, 0.90, 0.91, 0.81), where the two shape-parameter estimates apply to the two groups combined and the other three estimates apply to the control group alone. We refer to and treat these five estimates as the “true” values. We calculated z_W for all $2^5 = 32$ possible combinations of hypothesized and true parameter values, to enable us to assess how results could be affected by the misjudgment inherent in the hypothesized-versus-true discrepancy. Besides these 32 z_W -values, we found the (single) values of z_{LRNF} and z_{LRF} .

We obtained those 34 test statistics based not only on the cutoff date that actually occurred but also on three other cutoff dates 1, 2, and 3 years earlier. Thus, we used cutoff shortening times of $s = 0, 365, 730, 1095$ days and adjusted patients' values of q accordingly; $s = 0$ corresponds to March 24, 2009. Results that would have occurred with earlier cutoff than the actual one provide added perspective, as they may be less clouded by changes of therapy for control patients who experience progression.

Results for the test values, for all three tests and all four values of s , appear in Table 5. Not shown in the table are the findings from analyses of variance that we ran, separately for each s , on the 32 z_W -values. Generally, three-factor interactions among the five parameters were not statistically significant, and two-factor interactions were either nonsignificant or were significant statistically but not practically. Most main effects were strongly significant.

All four z_{LRNF} 's in Table 5 are > 4 and are far higher than any of the z_W 's or z_{LRF} 's. This is not a surprise, since Rini et al.¹² reported that progression-free survival differed between the two groups at a significance level < 0.0001 . As already noted, however, progression-free survival was predefined as a secondary, not the primary, endpoint. [The study found no difference at the 0.05 level (two-sided) for mortality, which was the primary endpoint.¹³]

Beyond the results for z_{LRNF} , the comparison of z_W versus z_{LRF} paints a different picture, and modestly favors z_W . For each of the four values of s , all 32 z_W -values are greater than the corresponding value of z_{LRF} .

Variability among the 32 values of z_W increases as s increases. This relation would seem to be due to the smaller number of deaths that arise with larger s (i.e., earlier cutoff entails fewer actual death times and more censored times).

Use of the hypothetical rather than the true values of the five parameters is not generally damaging for z_W . In fact, except for $s = 1095$, z_W -values are slightly higher with the hypothesized values. For $s = 0$, for instance, z_W is 1.60 if based on the hypothesized values (143, 289, 1, 1, 0.875) but is 1.54 for the true values (196, 402, 0.90, 0.91, 0.81).

The last five columns of Table 5 show the effects on z_W generated by the five parameters individually. For $s = 0$ for G , for example, the difference 0.02 is equal to $1.58 - 1.56$, where 1.56 is the mean of all 32 z_W 's and 1.58 is the mean of those 16 z_W -values that all have the hypothesized value 289 for G . By and large, as the table indicates, it is the true values of A and b but the hypothetical values of G , h , and p that yield the higher z_W 's. (Of course, a different clinical trial could have a distinction other than between A , b and G , h , p .)

This example has several interesting features. One log-rank test, LRNF, fares much better than W, though LRNF relates to a secondary rather than the pre-designated primary endpoint. The other log-rank test, LRF, does relate to the primary endpoint but yields z_{LRF} -values moderately but consistently below z_W (though even z_W is a bit lower than 1.96 except for $s = 1095$). The (inevitable) differences of the reasonable hypothesized parameter values from the true values do not hurt W overall. An anomaly is that z_W decreases, rather than increases, as time until data cutoff lengthens, apparently reflecting greater treatment-versus-control differences for progression than for mortality. The explanation could be that the treatment delays progression without the same appreciable delay for death, or (perhaps more likely) the control patients' option of switching therapy after they progress may prolong their survival beyond what it would be without that option.

8. General comments

The values of the five parameters (A, G, b, h, p) required by our test are hypothetical ones that pertain to the control distributions and are to be based on anticipations of what those distributions will be, perhaps reflecting historical data if available. The values must be specified before the trial starts, at the design stage (and should be included in the statistical section in the protocol). They could be selected in conjunction with power calculations. Those calculations, in turn, will be done along with the designing of the trial, which can consider costs as well as the relative roles and impacts of nonfatal and fatal events. The process of designing the trial, if it leads to our type of compromise test where neither event type plays the dominating role, may provide insights as to good parameter choices. (Of course, the complex design process also involves sample size, number of events needed, and trial duration. Power may be evaluated through simulation or other means. Required number of events could, e.g., be based on a linear combination of number of deaths and number of surviving patients with nonfatal events.)

Although generally parameter values should be chosen based on anticipations, Table 3 suggests possible small deviations from this rule for two of the parameters: It seems better for A to be high (relative to the true value, A^*) than low, and better for p to be low (relative to p^*) than high, because for A^* and p^* the last two table columns are monotonic in the same direction.

Although both $f_1(r)$ of (1) and $f_3(t|r)$ of (2) are Weibull distributions, there is no reason in principle that would prevent the use of other hypothetical distributions instead. Calculations would generally be more complex and power would be different, but the null distribution and Type I error would still be correct.

Our test disregards any nonfatal events after the first one and even disregards the first one if death is later observed. Although in principle it should be possible to harness the information from these disregarded nonfatal events to achieve greater power, that would introduce much greater complexity. Type I error is, of course, still correct despite the disregarded events. Our test already provides favorable power as evidenced by the simulation results.

In our test we make no provision for possible use of baseline covariates. Such an extension, if it could be done, would be beneficial.

Our test simply tests the null hypothesis that the joint distribution of time until the first (fatal or nonfatal) event and time until death is the same for the treatment and control groups. If the null hypothesis is true, that implies that the (marginal) distribution of the time until the first event is the same for the two groups, as is the (marginal) distribution of the time until death. By itself, rejection of the null hypothesis does not indicate whether the detected difference between the two groups pertains to time until first event, mortality, or both. This outcome is appropriate for the situation addressed in this paper, where, in advance of the trial, neither endpoint appears to be optimal. One should not use our test if only mortality or only a nonfatal endpoint is of interest; both nonfatal and fatal events must be of importance for the test to be appropriate. If the test rejects the null hypothesis, one may be able to get an idea of which of the two plays the greater role by examining separately the results for each.

Certain matters apply not just to our test but also more generally. In oncology, if progression occurs in a control patient, then the trial procedure may allow for switching the patient to the treatment arm or other therapies. With such a provision, aims are somewhat altered. The trial could then be effectively testing treatment right away versus treatment later, rather than treatment versus control. (Our example in Section 7 was somewhat of this nature.)

The treatment may be better than the control for the nonfatal endpoint but no different with respect to death. The log-rank test based on the nonfatal endpoint, but not the one based on death, could detect the difference. The former log-rank test would have a better chance of being significant than our test, whose power would be lower the greater the proportion of deaths observed. Even though the treatment does not delay death, it might (or might not) still result in better quality of life prior to death.

A more extreme situation, one that would seem less common but was mentioned by (e.g.) Finkelstein and Schoenfeld^{14, sec. 5}, would entail a treatment that is better than the control for the nonfatal endpoint but worse with respect to death (or vice versa). Such a condition poses a dilemma not only for our test but also for the two log-rank tests (which could give results opposite to each other). But presumably one could resolve the quandary, if it is recognized beforehand, by settling on one of the two endpoints (most likely death) as primary. Generally, our test should not knowingly be proposed at all for this case. But if it were used anyway, then our test might favor either arm (treatment or control) or neither, and results for the nonfatal and fatal endpoints would need to be examined separately.

We recognize, though, that our test may be problematic for a trial with (e.g.) a new drug that delays progression or improves symptoms, but hastens death. Before a trial, an investigator may sometimes have little idea whether such a perverse condition exists. Our test is generally appropriate in cases where the investigator is willing to rule out the (probably uncommon) condition where one arm is better for one endpoint and the other arm is better for the other. The present paper makes no attempt to duly address this condition or its ramifications. Our test is suitable, however, if the investigator is willing to accept either endpoint in itself (or a combination of the two) as evidence of benefit.

Our simulations used fixed numbers of patients, a fixed enrollment period, and a fixed time for data cutoff. In practice, a trial might operate somewhat differently, perhaps by stopping after a pre-specified number of deaths have occurred with little if any regard for number of patients randomized and without a fixed schedule for enrollment or cutoff. These operational differences should not affect the conclusions from our simulations.

The results of our simulations are largely favorable to our test. Although we used many Weibull conditions for our runs, other conditions, either Weibull or non-Weibull, could have led to results less favorable or more favorable. Many other conditions are possible. Thus, one could simulate using various combinations involving non-Weibull distributions for one or both of the true $f_1(r)$ and $f_3(t|r)$ and/or perhaps for one or both of the hypothetical $f_1(r)$ and $f_3(t|r)$. Our simulations tried to exploit the versatility of the Weibull distribution so as to represent a large variety of ways of choosing the true distributions in conjunction with the hypothetical ones, and, in fact, the 243 conditions in each of Tables 1 and 2 include 192 in which three or more of the five true parameters differ from the hypothetical ones.

The Wilcoxon test that we use, based on time or hypothetical expected time until death, bears only limited resemblance to the generalized Wilcoxon test of Gehan.¹⁵ The most obvious difference is that the latter test recognizes just one type of endpoint. In addition, that test effectively tallies ties between a control patient and a treatment patient in many cases where our test does not (i.e., our test breaks the tie).

Finkelstein and Schoenfeld¹⁴ proposed a test that they referred to as a generalization of Gehan's test. Where needed (i.e., in some instances where death or other failure is censored), that test breaks ties in the Gehan test by using a secondary measure such as infections or quality of life, evaluated at the lesser follow-up time of two patients.

In a trial for a rare disease with a limited patient pool and number of deaths, one could use the Finkelstein-Schoenfeld test to try to enhance power. Our test could also be suitable for such a trial, since it could likewise provide extra opportunity to demonstrate an effect.

Simulations run by Broglio and Berry,¹⁶ using a setting (but not tests) somewhat similar to ours, had a median time from randomization to progression that was greater for treatment than control patients but a median for post-progression survival that was the same for both. The results showed that, if the latter median was high, significance levels for a test with death as the endpoint were much poorer than for a test with nonfatal (progression) endpoint.

9. Concluding remarks

For a control-versus-treatment Phase III trial where it is not clear in advance whether mortality or survival free of nonfatal events provides a more suitable basis for a primary endpoint, this paper presents a simple test that can be a good compromise choice. Whereas the log-rank test based on a nonfatal event completely disregards death preceded by a nonfatal event and the log-rank test based on death fully disregards nonfatal events, our test totally disregards neither and thus offers an in-between strategy. In effect, we eschew both of the standard endpoints, the nonfatal one and death, and use instead an endpoint defined (in Section 2) by Y .

In oncology, there is ongoing debate as to which of the two usual endpoints is more appropriate. We note that, by itself, a nonfatal endpoint (such as progression-free survival), although a convenient and objective intermediate endpoint, has encountered regulators' questions as to its appropriateness in some cancers such as prostate and renal (Halabi et al.^{17, 18}). Recent regulatory approvals of novel drugs for treating men with advanced prostate cancer were based on trials that used co-primary nonfatal and fatal endpoints (Ryan et al.,¹⁹ Beer et al.²⁰). These trials^{19–20} were designed with co-primary endpoints where the Type I error rate was allocated to each of the endpoints that were monitored throughout the trial. It is noteworthy that, while the design of co-primary endpoints is feasible,²¹ the monitoring may be logistically more complicated due to the timing of the different events.

The new test applies the ordinary two-sample Wilcoxon statistic to patient values that consist of time from randomization until death if a patient has died, or time since randomization plus an increment otherwise. The calculations of the increments are based on five pre-specified parameters. One is the (anticipated) probability, p , that the first event is nonfatal. The others involve two Weibull distributions defined by (median, shape parameter) = (A, b) and (G, h) . If both Weibull shape parameters are 1 (i.e., both distributions are exponential), our test takes a particularly easy form, because then only three parameters remain to be (pre-)specified: the two medians (A and G) and p . For this case, for a patient who does not die before cutoff, the test increments time since randomization by a constant (same for every patient), equal to $G/(\log 2)$ if a nonfatal event has occurred or to the (usually larger) value $(A + pG)/(\log 2)$ otherwise.

We emphasize that, although the example in Section 7 pertains to cancer, our test can be used for other diseases also. Other possible applications are to cardiology as well as (e.g.) HIV/AIDS and cytomegalovirus disease.

We have presented a test that is an appropriate compromise if, before the trial, the investigator is unsure whether to use an endpoint tied only to fatal or more to nonfatal events. The test has correct Type I error, has good power, and is easy to use. We anticipate that our approach will be valuable and could (e.g.) see wide use in future Phase III trials where otherwise the investigator would be considering use of co-primary endpoints.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Stephen L. George for his many comments and suggestions. We also thank two reviewers and the Associate Editor for their comments. This work was supported in part by grants CA142538 and CA155296 from the National Cancer Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the NCI.

References

1. Borm GF, Teerenstra S, Zielhuis GA. Objective and perspective determine the choice of composite endpoint. *Journal of Clinical Epidemiology*. 2008; 61:99–101. [PubMed: 18177781]

2. Chi GYH. Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology*. 2005; 19:609–619. [PubMed: 16313272]
3. Lynch SE, Lavin PT, Genco RJ, et al. New composite endpoints to assess efficacy in periodontal therapy clinical trials. *Journal of Periodontology*. 2006; 77:1314–1322. [PubMed: 16881800]
4. Quan H, Zhang D, Zhang J, et al. Analysis of a binary composite endpoint with missing data in components. *Statistics in Medicine*. 2007; 26:4703–4718. [PubMed: 17431851]
5. Sandrini G, Dahlöf CG, Mathew N, et al. Focus on trial endpoints of clinical relevance and the use of almotriptan for the acute treatment of migraine. *International Journal of Clinical Practice*. 2005; 59:1356–1365. [PubMed: 16236092]
6. Song R, Cook TD, Kosorok MR. What we want versus what we can get: A closer look at failure time endpoints for cardiovascular studies. *Journal of Biopharmaceutical Statistics*. 2008; 18:370–381. [PubMed: 18327727]
7. Wong WK, Furst DE, Clements PJ, et al. Assessing disease progression using a composite endpoint. *Statistical Methods in Medical Research*. 2007; 16:31–49. [PubMed: 17338293]
8. Hougaard, P. *Analysis of Multivariate Survival Data*. New York: Springer; 2000.
9. Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2. Hoboken, NJ: Wiley; 2002.
10. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*. 2007; 26:2389–2430. [PubMed: 17031868]
11. Lehmann, EL. *Nonparametrics: Statistical Methods Based on Ranks*. Vol. 2. San Francisco: Holden-Day; 1975.
12. Rini BI, Halabi S, Rosenberg JE, et al. Bevacizumab plus interferon alpha compared with interferon alpha monotherapy in patients with metastatic renal cell carcinoma: CALGB 90206. *Journal of Clinical Oncology*. 2008; 26:5422–5428. [PubMed: 18936475]
13. Rini BI, Halabi S, Rosenberg JE, et al. Phase III trial of bevacizumab plus interferon alpha versus interferon alpha monotherapy in patients with metastatic renal cell carcinoma: Final results of CALGB 90206. *Journal of Clinical Oncology*. 2010; 28:2137–2143. [PubMed: 20368558]
14. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*. 1999; 18:1341–1354. [PubMed: 10399200]
15. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. 1965; 52:203–223. [PubMed: 14341275]
16. Broglio KR, Berry D. Detecting an overall survival benefit that is derived from progression-free survival. *Journal of the National Cancer Institute*. 2009; 101:1642–1649. [PubMed: 19903805]
17. Halabi S, Vogelzang NJ, Ou S-S, et al. Progression-free survival as a predictor of overall survival in men with castrate-resistant prostate cancer. *Journal of Clinical Oncology*. 2009; 27:2766–2771. [PubMed: 19380448]
18. Halabi S, Rini B, Escudier B, et al. Progression-free survival as a surrogate endpoint of overall survival in patients with metastatic renal cell carcinoma. *Cancer*. 2014; 120:52–60. [PubMed: 24347384]
19. Ryan CJ, Smith MR, de Bono JS, et al. Abiraterone in metastatic prostate cancer without previous chemotherapy. *New England Journal of Medicine*. 2013; 368:138–148. [PubMed: 23228172]
20. Beer TM, Armstrong AJ, Rathkopf DE, et al. Enzalutamide in metastatic prostate cancer before chemotherapy. *New England Journal of Medicine*. 2014; 371:424–433. [PubMed: 24881730]
21. Halabi S. Adjustment on the type I error rate for a clinical trial monitoring for both intermediate and primary endpoints. *Journal of Biometrics and Biostatistics*. 2012; S7:015.10.4172/2155-6180.S7-015

Table 1

Percentage of 1000 runs in which $z_W > z_{LRNF}$, by A^* , G^* , b^* , h^* , p^* .

$b^*=b^{**}$	$h^*=h^{**}$	p^*	$A^* = 2.4, G^* =$			$A^* = 3.0, G^* =$			$A^* = 3.6, G^* =$		
			2.4	3.0	3.6	2.4	3.0	3.6	2.4	3.0	3.6
7/8	7/8	0.65	64.9	63.6	59.2	68.4	63.5	59.2	62.7	61.4	59.9
		0.75	74.2	68.5	64.8	72.1	66.8	62.7	69.2	64.1	62.5
		0.85	76.5	71.4	67.3	76.0	69.3	64.8	74.4	67.0	64.2
1	1	0.65	71.7	64.7	64.9	70.0	65.5	64.1	67.9	62.2	61.0
		0.75	78.1	72.3	65.1	73.3	71.2	66.6	73.1	67.0	63.9
		0.85	83.3	77.4	72.9	82.0	74.7	72.4	76.8	74.4	67.5
8/7	8/7	0.65	78.3	71.5	69.7	76.6	70.3	64.4	72.9	68.6	63.1
		0.75	84.2	76.1	71.1	80.9	75.7	72.2	77.1	71.0	68.1
		0.85	87.7	84.1	77.4	85.8	79.0	75.1	79.3	76.7	70.8
1	7/8	0.65	63.3	57.1	52.3	65.4	57.9	58.0	63.5	57.8	55.9
		0.75	67.3	63.6	60.8	68.4	64.9	57.4	66.1	62.1	57.2
		0.85	70.3	65.9	59.8	70.2	65.7	60.5	70.4	63.0	61.2
1	1	0.65	68.5	64.3	57.0	70.0	61.9	61.7	67.7	60.8	56.3
		0.75	74.2	67.5	65.0	73.8	68.1	64.7	70.6	64.3	61.8
		0.85	80.7	74.4	69.1	78.7	74.3	69.3	71.8	72.2	67.6
8/7	8/7	0.65	73.6	64.8	62.8	73.7	68.6	63.2	70.4	65.9	62.2
		0.75	77.8	73.2	68.2	77.6	71.8	67.3	74.7	69.8	67.5
		0.85	85.1	77.5	73.0	83.6	76.8	69.0	77.4	72.7	69.3
8/7	7/8	0.65	58.5	49.6	47.8	60.6	57.5	51.4	58.8	57.3	52.2
		0.75	60.1	55.4	52.7	65.6	60.2	55.7	63.1	58.8	56.6
		0.85	67.4	63.3	56.9	72.1	63.6	61.5	65.8	66.1	58.7
1	1	0.65	60.7	55.7	50.5	65.1	59.1	55.1	64.1	56.9	54.4
		0.75	68.2	61.3	54.5	69.0	63.4	59.3	69.5	62.6	58.1
		0.85	73.3	65.0	62.0	74.6	68.2	60.6	72.0	66.2	63.3
8/7	8/7	0.65	65.2	59.3	53.1	68.3	63.3	60.4	66.8	60.7	57.3
		0.75	73.6	65.7	59.5	73.7	67.4	61.3	68.2	63.9	61.9
		0.85	81.0	72.0	67.9	80.3	76.2	66.7	77.1	71.7	66.3

Table 2

Percentage of 1000 runs in which $z_W > z_{LRF}$, by A^* , G^* , $b^* = b^{**}$, $h^* = h^{**}$, p^* .

$b^* = b^{**}$	$h^* = h^{**}$	p^*	$A^* = 2.4, G^* =$			$A^* = 3.0, G^* =$			$A^* = 3.6, G^* =$		
			2.4	3.0	3.6	2.4	3.0	3.6	2.4	3.0	3.6
7/8	7/8	0.65	64.4	64.4	72.9	58.0	61.6	67.6	54.3	58.5	61.3
		0.75	61.8	67.4	70.5	56.9	61.4	65.2	53.2	54.5	59.3
		0.85	61.6	67.2	68.7	58.4	60.9	64.2	49.9	58.2	59.5
	1	0.65	59.2	66.1	69.1	55.4	61.7	65.1	50.2	55.0	59.3
		0.75	62.6	60.5	69.8	56.8	59.0	64.1	51.0	54.9	59.1
		0.85	57.0	64.6	66.3	52.1	57.0	62.6	51.6	50.9	59.1
	8/7	0.65	53.2	61.3	64.2	48.7	55.8	62.7	44.2	51.8	55.5
		0.75	53.5	60.1	64.8	48.5	55.7	59.9	42.2	54.0	55.8
		0.85	53.8	58.6	66.8	47.7	52.8	59.9	46.1	46.7	53.7
1	7/8	0.65	71.7	77.7	81.0	67.2	71.9	74.6	62.3	64.8	68.1
		0.75	74.2	76.1	78.1	66.6	71.2	73.8	61.7	63.7	66.3
		0.85	72.2	76.7	79.1	68.6	70.0	75.2	59.9	63.6	71.1
	1	0.65	68.3	72.9	79.2	62.1	68.6	71.2	57.0	63.6	67.1
		0.75	69.7	74.5	77.5	64.3	69.3	69.9	57.0	62.5	66.4
		0.85	68.9	72.0	77.1	63.8	64.2	68.2	57.7	59.6	62.1
8/7	7/8	0.65	59.9	72.6	75.1	58.3	65.8	69.5	51.8	61.3	64.7
		0.75	66.8	71.2	76.5	60.9	65.2	67.4	52.8	59.9	62.6
		0.85	65.3	71.4	74.5	57.9	66.2	69.3	54.1	60.2	61.9
	1	0.65	79.3	84.0	87.4	75.7	78.6	83.3	70.3	73.3	76.5
		0.75	81.7	85.9	86.4	78.5	79.8	83.3	69.6	72.9	75.3
		0.85	81.8	83.0	86.6	77.0	77.1	80.5	69.7	74.4	75.1
8/7	7/8	0.65	77.4	83.2	85.7	71.5	80.0	80.4	66.1	71.4	73.9
		0.75	79.4	83.5	85.8	71.9	79.7	81.1	65.8	72.1	72.9
		0.85	79.3	82.6	83.0	74.6	78.6	80.9	65.2	71.1	74.6
	8/7	0.65	75.5	81.2	84.7	71.2	75.7	78.0	64.8	69.3	71.8
		0.75	74.4	80.7	85.7	70.4	76.5	80.6	62.8	71.1	73.5
		0.85	74.8	79.2	82.9	70.3	74.3	76.8	62.1	66.5	69.9

Table 3Main effects [averages of $(z_W - z_{LRNF})$ and $(z_W - z_{LRF})$] at each level of each of the five true parameters

Parameter	Level	Average of	
		$(z_W - z_{LRNF})$	$(z_W - z_{LRF})$
$A^* =$	2.4	0.262	0.287
	3.0	0.238	0.236
	3.6	0.192	0.172
$G^* =$	2.4	0.292	0.155
	3.0	0.229	0.236
	3.6	0.171	0.304
$b^* = b^{**} =$	7/8	0.283	0.108
	1	0.235	0.226
	8/7	0.174	0.361
$h^* = h^{**} =$	7/8	0.169	0.270
	1	0.229	0.235
	8/7	0.294	0.190
$p^* =$	0.65	0.161	0.215
	0.75	0.227	0.236
	0.85	0.304	0.244

Table 4

Percentage of outcomes in which each test, based on its value of z , is worst (smallest value of z), middle, and best among the three tests

Test	Worst	Middle	Best	Total
W, based on z_W	4.1%	57.2	38.7	100%
LRNF, based on z_{LRNF}	48.0	21.1	30.9	100
LRF, based on z_{LRF}	47.8	21.7	30.5	100
Total	100	100	100	

Table 5

For CALGB 90206 data, values of z_W (summarizations for 32 z_W 's), z_{LRNF} , and z_{LRF} , for the actual cutoff date and 1, 2, 3 years earlier ($s = 0, 365, 730, 1095$ days earlier)

			32 z _W -values		z _W -value if all 5 parameters have the values that are		Difference above (+) or below (−) the mean of all 32 z _W -values, for the mean of those 16 z _W 's that use the hypothesized value for				
s	Test	z _{LRNF} or z _{LRF}	Mean	Range	Hypothesized	True	A	G	b	h	p
0	W		1.56	1.52 to 1.60	1.60	1.54	0.00	0.02	−0.00	0.02	−0.00
0	LRNF	4.31									
0	LRF	1.46									
365	W		1.65	1.57 to 1.72	1.71	1.62	−0.00	0.02	−0.00	0.03	0.00
365	LRNF	4.35									
365	LRF	1.51									
730	W		1.78	1.59 to 1.88	1.83	1.77	−0.03	0.02	−0.03	0.05	0.01
730	LRNF	4.57									
730	LRF	1.39									
1095	W		2.13	1.60 to 2.52	2.11	2.16	−0.14	0.05	−0.11	0.11	0.06
1095	LRNF	5.01									
1095	LRF	1.33									