Published in final edited form as:

Stat Med. 2011 December 30; 30(30): 3507–3519. doi:10.1002/sim.4398.

A hierarchical rank test for crossover trials with censored data

Erica Brittain and Dean Follmann

National Institute of Allergy and Infectious Diseases, Bethesda, MD

Abstract

We propose an approach to analyze survival time in a crossover clinical trial by performing a primary ranking on whether events occur and a secondary ranking on event times. This hierarchical ranking method (HR) is meant to reflect the idea that the goal of therapy is to prevent a clinical event, and failing that, to delay the occurrence of the event, hopefully for a substantial amount of time. We compare our approach to other methods including one method proposed by Feingold and Gillespie, a recommended procedure. The power is similar in many settings but the hierarchical ranking can have substantially greater power under certain censoring patterns and also under a cure model, or models where treatment induces a substantial delay in some fraction of patients. We additionally feel that the HR method should be more clinically relevant in many settings. The method can also be applied to continuous outcomes censored by a limit of detection, such as HIV viremia.

Keywords

Crossover Trials; Cox Regression; Subjective Rankings; Survival Times

1 Introduction

Crossover trials are an important and efficient design to evaluate quick acting interventions whose effects disappear quickly upon cessation. Under the simple two-treatment two-period crossover design, patients are randomized to one of two sequences, e.g. AB or BA where an initial treatment is administered for a fixed period of time and an endpoint is recorded. Patients then receive the opposite treatment and the process repeated. Frequently an interlude or “washout” separates the two periods to mitigate any dissipating effect of treatment on the second period outcome. For a continuous endpoint such as a blood pressure, analysis strategies are well developed (See Senn [1] for comprehensive discussion) and are based on calculating the difference in blood pressure between the two treatment periods. As each patient effectively serves as his own control, this design can be much more efficient than a standard parallel design where each patient only receives a single treatment.

In some settings the endpoint of interest is the occurrence or timing of an event. Consider an anti-angina medication where treatment is evaluated at the end of study by recording the time to > 1 mm depression of the ST segment of an ECG (electrocardiogram). This time may be censored as not all patients will have such an event. A crossover trial would be justified if the anti-angina medication does not linger and does not impact the underlying disease. A second example involves the effect of a monoclonal antibody on the occurrence of anaphylactic shock in patients with recurrent episodes, as monoclonal antibodies dissipate

relatively quickly one might consider a crossover approach evaluating the time to the first event in each treatment phase.

Analysis for such a crossover design with potentially censored failure times have been addressed by other authors. France et al. [2] proposed treating each individual as a stratum in Cox regression, effectively assessing whether there is any treatment preference within the individual. Feingold and Gillespie [3] proposed several tests in this setting, one of which is based on a generalized Wilcoxon approach (we denote this procedure FG-W); Senn [1] in a crossover review paper recommends the FG-W method for analyzing survival data in crossover studies. This approach takes the difference in event times between periods and uses a Wilcoxon test (generalized to both left- and right-censored data) to compare sequence groups. As in the Gehan formulation of the Wilcoxon test, some pairs of patients may not be able to be ordered; and in fact patients with no events in either period are discarded from the analysis. Feingold and Gillespie find that their FG-W approach has more power than the Cox-based test of France et al. [2], analogous to the difference between a signed rank test and a sign test.

In this paper, we propose an alternate procedure for analysis, which we term Hierarchical Ranking (HR), which flows from previous consideration of subjective ranking (see e.g., Follmann et al [4], Brittain et al [5], Neaton et al [6], Bjorling & Hodges [7]). The basic idea is that avoidance of an event during a fixed interval is much more clinically meaningful than delaying an event within that interval and the ranking of patients should reflect this hierarchy. Consequently, our first ordering is based on whether the occurrence of events differs. The second ordering is based on the relative times of the events, so that time differences are effectively only tie-breakers. That is, patients with a single event get the most extreme ranks; those with an event in both periods, or no events at all will be in the middle.

In Section 2, we introduce notation and an example data set. Sections 3 and 4 review existing procedures for treatment preference in crossovers with censored outcomes. We motivate and describe HR, and illustrate in general terms how it compares to existing methods in Section 5. Simulation results that compare methods under a wide range of models are presented in Section 6. Section 7 presents an example; which is followed by a discussion section.

2 Two-treatment two-period crossover trial

We assume that patients are randomized to one of two treatment sequences, say A followed by B (AB) or BA. Patients receive the treatment and are followed until the event occurs or a fixed period of time say τ , whichever occurs first, so that there is constant censoring time across patients. After any suitable washout period, patients receive the opposite treatment and are again followed until the event or a period of time τ has passed. Let T_{ij} be the failure time for patient $i = 1, \dots, n$ in period $j = 1, 2$. We let Z_{ij} be the indicator of treatment A for person i in period j , thus $(Z_{i1}, Z_{i2}) = (1, 0)$ or $(0, 1)$.

We define the counts $N_{k\ell}$ as the number of patients with $k = 0, 1$ events in Period 1 and $\ell = 0, 1$ events in Period 2. Thus N_{11} is the number of patients with events in each period. We

further define N_{11}^1 as the number of patients with $T_1 > T_2$ and who thus “prefer” Period 1. N_{11}^2 is similar. When needed, we will use the notation $N_{k\ell}(AB)$ to denote the value of $N_{k\ell}$ of those patients randomized to sequence AB. Note that $N_{11} = N_{11}^1 + N_{11}^2$ and that $N_{k\ell} = N_{k\ell}(AB) + N_{k\ell}(BA)$.

To help fix ideas, consider the data of Table 1 for a 2 period crossover trial with $\tau = 1$ where 6 patients are randomized to AB and 9 randomized to BA. Patient 1 is censored at $\tau = 1$ on B in Period 1 and has an event at time=.1 in Period 2 on A. Patient 9 is censored in both periods, while Patient 10 has $T_2 - T_1 = .45$ and thus prefers Period 2 to Period 1 as the event is avoided longer.

We display the numbers of events for these data in Figure 1, a 2×2 table of event occurrence by period, augmented by classifying patients with 2 events by whether $T_1 > T_2$ or vice versa, using N_{11}^1 and N_{11}^2 , respectively.

3 Events based procedures

A number of simple methods for analyzing crossover trial data with paired binary outcomes have been proposed. While intuitively use of the failure time provides additional information and will be incorporated later, it is of some interest to review these simpler methods. Our discussion is largely based on the textbooks by Jones & Kenward [8] and Senn [9].

The simplest approach is probably McNemar’s test, which classifies patients in terms of treatment preference. The number of patients who prefer treatment A to B is $N(A \succ B) = N_{01}(AB) + N_{10}(BA)$ where preference is defined as the avoidance of an event. Under the null hypothesis of no preference and assuming no period effect, $N(A \succ B)$ has a binomial distribution with probability 1/2 and number of “trials” $N(A \succ B) + N(A \prec B)$. Eight patients from Table 1 have discordant values between the periods, and all eight favor B, so the p-value is .0078. A problem with this test, as implied by Prescott [10] is that the type I error rate can be inflated when unequal randomization is coupled with a period effect. For example, consider the case where A and B have equal efficacy, more patients are allocated to AB than to BA, and where patients do better in Period 2 than Period 1, then B will tend to look better than A with McNemar’s test.

The Mainland-Gart test improves on McNemar’s such that the Type I error problem under a period effect is avoided. A key consideration in crossover trials is the potential for period effects, and the framework of a two sample test comparing AB to BA on Period 1 versus Period 2 differences appropriately addresses this potential. In Mainland-Gart, a 2 by 2 table, that compares Period 1 preference to Period 2 preference by randomization sequence AB and BA can be constructed. If A is superior to B, this will lead to more preference for Period 1 under AB and more preference for Period 2 under BA. Testing of the independence of rows and columns (essentially a test of treatment, adjusted for period) can be accomplished by Fisher’s exact test or a chi-square test; that is an analysis of a 2 by 2 subtable of Table 2 where the middle column “Neither” is eliminated. (When analyzing Table 1 data, the AB arm has 0 patients with Period 1 better and 3 with Period 2 better, whereas BA has 5 with Period 1 better and 0 with Period 2 better, leading to a p-value of .0179.)

Both of the above tests ignore the patients with events in both periods or patients with no events at all. Prescott [10] introduced a test incorporating the tied patients that corresponds to a permutation t-test where we assign the scores $-1, 0, 1$ to patients with preference for the first period, no preference, or preference for the second period, respectively (see Table 2). It can also be viewed as an example of the Cochran-Armitage Trend test. Prescott conducts simulations which indicate that Prescott's test has better power than the Mainland-Gart with or without a period effect and has better power than McNemar's test in the presence of a period effect. Prescott provides a rationale for the improved power of his test compared to excluding the doubly censored patients; his reasoning involves random effects models. Another way to understand this is to note that Prescott compares AB to BA on a three point ordered scale: Period 1 preference, Period 1 and Period 2 tied, Period 2 preference; that is, this is a continuum where the tied values are intermediate. If A is better than B, this will lead to AB having a distribution that is shifted towards the Period 1 preference end of the scale, more than BA's distribution. Just as one would not exclude the middle values of any ordered three point scale, including observations seen at the middle level of this three point scale improves power, as demonstrated by Prescott. (When the Prescott procedure is applied to Table 1 data, the AB arm had 0 with Period 1 better, 3 tied, and 3 with Period 2 better; whereas BA has 5 with Period 1 better, 4 tied, and 0 with Period 2 better, resulting in a p-value of .0084.)

It is straightforward to consider an extension of all these event based procedures, where patients with events in both periods prefer the treatment with the later event time, as illustrated in Section 2 and Figure 1. We evaluated such an extension to Prescott, which would modify the layout presented in Table 2 by moving N_{11} patients who preferred Period 1 from the tied category into Period 1 preference category, and likewise for those who preferred Period 2. However, in subsequent simulations (not shown), the extended version surprisingly performed more poorly than the regular Prescott in most scenarios. It appears that the extended Prescott yielded a larger treatment difference, but bigger standard deviation, often leading to a smaller non-centrality parameter. However, there are potentially scenarios when the extended version of all these event based procedures would have improved power.

4 Time to event procedures

France, Lewis and Kay [2] introduced a set of methods based on patient preference using failure times. Here, a patient prefers treatment A to B if the failure time on A exceeds B and France et al suggest using a Cox model for analysis where patient i has hazard function

$$\lambda_i(t) = \lambda_{0i}(t) \exp(x_1 \beta_1 + x_2 \beta_2)$$

where x_1 identifies period and x_2 identifies treatment. Thus each patient defines a stratum and patients with no events in either period do not contribute to the partial likelihood. Note that this method does not take into account the magnitude of the difference in failure times, just whether a preference is identified and the method ends up being very similar to the Mainland-Gart procedure.

Feingold and Gillespie [3] proposed a different method of analysis motivated by an extension of the Wilcoxon statistic to both left- and right-censored data (Gehan [11], Mantel [12]). They form the difference in event times $T_{i2} - T_{i1}$ which may be left or right censored, or completely undefined if censoring occurs in both periods. They then use the Gehan method to determine a preference between pairs of patients, such that if no preference can be definitively established because of the particular censoring times involved, then $U_{ii'}$ is zero;

$$U_{ii'} = \begin{cases} 1 & \text{if } T_{i2} - T_{i1} > T_{i'2} - T_{i'1} \\ -1 & \text{if } T_{i2} - T_{i1} < T_{i'2} - T_{i'1} \\ 0 & \text{otherwise} \end{cases}$$

For patient i , the score U_{i+} is formed by summing over all patients i' . Note that if A is better than B then subjects in the AB group should tend to have +1s and subjects in the BA group should tend to have -1s. The sum of U_{i+} for Group 1 is the numerator of the test statistic, with permutation variance given by Gehan [11] and Mantel [12]. This test was developed for an additive model; that is, a shift in survival curves, rather than the usual multiplicative model assumed in survival analysis; nonetheless, the authors report better power than France et al, with data simulated under either model. Presumably the superior power is because this method of analysis does take into account the magnitude of the difference in failure times and is preferred to the France et al test by Senn [1]). As noted earlier, this is akin to preferring the signed-rank test to a sign test. (In the same paper Feingold and Gillespie [3] propose another method for this setting which computes difference of each rank for each subject, and computes a t-test between sequence groups. Unlike FG-W, this score transformation test does not discard doubly censored patients. However, Feingold and Gillespie [3] found this other method has less statistical power than FG-W, especially with greater heterogeneity between patients. We do not study this procedure.)

There is an extensive literature on paired survival data analysis, and Desu and Raghavarao [13] devote a section in their book to analyses of paired data with uniform follow-up time, which is very related to our setting. However, these methods, including Woolson and Lachenbruch's generalized signed rank test [14] are designed for traditional paired data, such as randomization within sets of twins, where members of the pair have no natural ordering, and this naturally leads to a one sample comparison of paired data, as there is no analogue of period effect here. This is different from the crossover design where we randomize to ordered AB or BA, and we make a two sample comparison of Period 1 to Period 2 differences to be robust to period effects.

5 Hierarchical ranking

Interventions are designed ideally to prevent a clinical event, and failing that, to delay occurrence of the event. To obtain a statistical method that reflects this ideal, we propose a hierarchical ranking procedure for time to event outcomes where patients are ranked in terms of the strength of this preference for a given arm. We also recognize that consideration of a statistical cure model would lead to similar ordering.

Our ranking is best introduced by example. Consider the sequence AB. If a patient avoids an event on B but not on A, then they prefer B to A. Additionally, if Patients 1 & 2 have events on A, but no events on B, and Patient 1's event time on A is earlier than that of Patient 2, then we feel that Patient 1 is demonstrating a stronger preference for B over A than is Patient 2. Since avoiding an event is so desirable, we feel that patients who have events in both periods are displaying a weaker preference than patients who avoid an event in one period. Put another way, avoidance of an event in one period suggests the actual difference of event times is likely to be greater than when both events are observed. This kind of thinking leads to the following hierarchical ranking of patient preference for Period 1 over Period 2. For simplicity, we assume continuous failure times.

- Patients who have a single event get the most extreme ranks. Suppose N_{01} patients have events in Period 2 but not Period 1. They receive the smallest ranks $1, \dots, N_{01}$ based on the timing of the event in Period 2. Conversely, the N_{10} patients with events in Period 1 receive the largest ranks $N - N_{10} + 1, N - N_{10} + 2, \dots, N$.
- Patients with events in each period receive less extreme ranks. The N_{11}^1 patients who have $T_1 > T_2$ and thus prefer Period 1, receive the ranks $N_{01} + 1, \dots, N_{01} + N_{11}^1$ based on $T_1 - T_2$. Conversely, the N_{11}^2 patients with $T_2 > T_1$ receive the ranks $N - N_{10} - N_{11}^2 + 1, \dots, N - N_{10}$.
- The N_{00} patients with no events in either period display an equal preference for Period 1 and Period 2 and thus each receive the average of the remaining ranks or

$$\frac{N_{01} + N_{11}^1 + 1 + N - (N_{10} + N_{11}^2)}{2}.$$

To help understand the ranking scheme it is instructive to consider Figure 2, which expands the display of data from Figure 1. The smallest (largest) ranks are in the Southeast (Northwest) quadrant. We see that hierarchical ranking can be viewed as a generalization of Prescott's test where the "tied" observations among patients with a single event are broken by the timing of the event. Additionally, the "tied" observations among patients with two events are broken by first the sign of the difference and then the difference in the event times.

If treatment A is more effective than B, then patients randomized to AB should tend to have a preference for Period 1 and smaller ranks. Conversely patients in the BA sequence should have a preference for Period 2 and larger ranks. Thus we use the two group (AB versus BA) Wilcoxon statistic to test for a treatment effect. For large trials, the usual normal approximation to the Wilcoxon statistic, allowing for ties, can be used; as shown in Lehmann [15]. For smaller trials a permutation approach, where the sequences, AB and BA are permuted and the resultant test statistic calculated, can be used to obtain an exact null distribution.

It is interesting to compare the FG-W test to HR under a variety of extreme settings. First, if there is no censoring so that $N = N_{11}$, HR corresponds exactly to FG-W. Second, suppose that all events in Period 2 are censored due to an overwhelming secular trend; for example,

if a crossover is done over two allergy seasons, where the first season has many events and the second season is much milder and has no events. In this setting FG-W has literally zero power as all patients are tied. HR, on the other hand, reduces to the usual two-group Wilcoxon statistic using just Period 1 data. Finally, suppose a superlative treatment (A) has extreme carryover where patients randomized to AB are all doubly censored and all patients randomized to BA have events in Period 1 and are censored in Period 2. Again, FG-W breaks down but HR reduces to the Fisher's Exact Test and also Prescott's test. HR reduces to the Fishers Exact because there is a complete separation in the ranking: all AB patients are tied and ranked above all BA patients, such that an exact Wilcoxon test yields the same p-value as the Fishers exact test. Thus HR seamlessly adapts to the proper test under different extreme scenarios and should do well for settings "close" to the extreme.

To explore the differential ranking of HR and FG-W in a nonextreme setting, consider the data of Table 1. The p-values are similar: .0076 for the HR and .0145 for FG-W. The patient who experienced an event at .95 in the first period and .10 in the second period was ranked 6 out of 15 under HR, but 3 out of 15 under FG-W. Under HR, there are five patients ranked before this observation because they had an event in Period 2 and not in Period 1. In contrast, under FG-W this patient, with a known event time difference of $-.85$ is considered more extreme than those with a late event in Period 2 and no event in Period 1. Such a patient might have a smaller difference in time than $.85$, if they were observed.

One setting where we much prefer the avoidance of an event on A relative to B during trial follow-up as opposed to delaying A relative to B during follow-up is when censoring indicates a high probability of cure. Suppose that treatment cures a certain fraction of patients so their event time is infinite, but has no effect on time to event for the remainder. Under such a cure model, ranking based on the timing of events is much less important than whether an event occurs or not. Similar statements apply if treatment moves a fraction of patients to have events after the censoring times. Under such models, Prescott's test should be optimal, although HR should be fairly close. Let S_{kl} denote the set of patients with k events in the first period and l events in the second; S_{11}^1 denotes the set where patients have events in both periods but with a longer survival in Period 1; S_{11}^2 is similar. Then HR essentially takes the tied observations within the S_{01} , S_{11}^1 , S_{11}^2 , S_{10} "sets" and breaks ties within each set at random, whereas Prescott's test essentially treats all observations within S_{01} as tied, and likewise for S_{10} and S_{11} . FG-W can be similar to HR, but its ranking procedure can lead to a patient in S_{01} being ranked as less extreme than a patient in S_{11}^1 . This occurs in the data of Table 1, where Patients 4 and 5 of S_{01} are ranked less extreme than Patient 6 of S_{11}^1 under FG-W.

While our motivation for the HR approach has focused on time-to-event endpoints, these arguments may also apply to data that are subject to censoring for other reasons such as being above or below the limit of detection, provided being censored is qualitatively different from the uncensored values, and that the limit of detection is constant across tested samples. In studies of HIV, an important outcome is whether a patient has a plasma viral load of less than 50 copies/ml, the limit of detection of many current assays. Achieving an outcome below the limit of detection is a goal of anti-retroviral therapy and patients who are

above the limit of detection are candidates for therapy change. Thus, achieving or not achieving 50 copies/ml has a substantial impact on patients. A related setting where the HR approach might be appealing is in crossover trials of malaria prophylaxis that prevents or quickly reduces parasitemia following an exposure. At the end of each period, parasitemia is measured and patients without parasitemia are considered uninfected which is the ideal goal of the prophylaxis, but less parasitemia is also desirable. Thus the HR ranking reflects the clinical importance of the malaria prophylaxis. A crossover trial here can work because at the end of the first period all patients could be treated with antimalarial drugs. These drugs can completely eradicate infection, but leave patients at the same risk of being infected at the start of the second period.

6 Simulations

To evaluate the performance of the different methods, we simulate two period crossover trials with survival outcomes under various scenarios. We generate failure times under a Gamma mixture of Weibulls. Patient i draws a baseline intensity λ_i from a Gamma distribution with mean μ and standard deviation σ . Conditioned on this λ_i , the time to failure has the following hazard:

$$\frac{kt^{k-1}}{[\lambda_i \exp\{Z(t)\beta\}]^k}$$

where k is the Weibull shape parameter and $Z(t)$ is the treatment indicator at time t , where t is the time since the start of the period. We let $k = 1$, which provides an exponential (constant) conditional hazard, $k = 3$ which provides an increasing conditional hazard, and $k = .5$ for a decreasing hazard. We set $\mu = E(\lambda_i) = .20$, let $\sigma = \sqrt{\text{Var}(\lambda_i)}$ range from .01 to .10, and set n per randomization sequence equal to 50.

We evaluate a Cox analysis using just first period data and then a crossover design using Prescott, the stratified Cox (France et al), FG-W, and HR analyses. The presented power uses the method described in the FG-W paper, which uses a normal approximation and excludes patients who are censored in both periods. Each set of parameters is studied with 10,000 replications.

The simulation results for the $n=50$ case are shown in Table 3. In all scenarios shown in Table 3, HR and FG-W have very similar power; both of these have much higher power than the other methods, except for the Prescott test, where the advantage in power is much more variable. In the parallel design, greater heterogeneity (i.e., larger σ) is associated with loss of power, particularly under the Weibull with increasing hazard; however, in the crossover design, there is no such decline with HR and FG-W, suggesting a crossover design is especially useful under heterogeneity, as expected. The results of Prescott's test were surprisingly good in a range of settings, except for the Weibull $k=3$ case. However, unlike HR and FG-W, the standard Prescott test would have very low power in situations where (nearly) everyone has events, a setting the simulations do not consider.

We note that similar patterns of all of the above results were observed for $n=10$. Rejection proportions under the null were computed for these and subsequently presented simulations in this paper, with 10,000 replications; the estimated Type I error for HR was very well controlled, with all values statistically consistent with .05. Similarly, the rejection proportions for all other procedures were very well controlled, except for the Cox based analyses which showed slight inflation for $n=50$, and a little more inflation for $n=10$. For brevity, these null results are not shown.

6.1 Special models

We also assess power under some special models. Except where noted otherwise, event times are generated under a Gamma mixtures of exponentials with mean λ of .20 and $\sigma=.05$. First, we consider the case where the event rate in the second period is reduced; this is examined over a range of factors. Second, we explore a carry-over effect model, where if a patient receives an active regimen in Period 1, .75 of the benefit in Period 1 carries over to the patient's second period. That is, a patient's parameter, λ_i is multiplied by e^β while assigned to the active treatment in the first period. This modification changes to $.75e^\beta$ in Period 2; whereas patients who start with placebo are generated according to λ_i in the first period and $\lambda_i e^\beta$ in the second period. In these first two models, the hazard ratio, e^β , is 1.83. Third, we simulate under a cure model, where patients have events according to their λ_i values throughout both periods, with the following exception. Patients who are assigned to the active treatment in Period 1 have a 50% probability of being "cured", and these patients have no events during either period. Patients who are first treated by placebo have a 50% probability of being "cured" in the second period when they receive the treatment, such that these patients have no events in Period 2. Again, the number of replications for each parameter set is 10,000, and there are 50 observations per group; Table 4 presents rejection rates. As expected, the HR has much higher power than the FG-W when the ratio of the Period 2 to Period 1 censorship factor is large. In addition, HR has slightly higher power than FG-W under carry-over. And, finally, with a cure model, the HR performs much better than the FG-W; however, Prescott and the single period Cox analysis are the top performers.

6.2 Example with recurrent events

A different kind of paired survival data comes from events that recur, where the risk of an event at a given moment might depend on how long ago the previous event occurs. For example, the risk of hospitalization for depression might be lowest immediately after hospital discharge. In this section we focus on a design for recurrent events that we encountered in a recent consultation. In this case, investigators were interested in the effect of a monoclonal antibody in patients with frequent episodes of spontaneous (idiopathic) anaphylaxis. Idiopathic anaphylaxis occurs in the absence of known allergies, such as to food or insect venoms; and where no other precipitating cause is identified. These episodes in part involve mast cells, which release chemicals that cause systemic symptoms and anaphylactic shock. The monoclonal antibody is hypothesized to reduce a critical activation receptor on the mast cells, thus preventing anaphylaxis. Monoclonal antibodies dissipate relatively quickly, so a crossover trial might be viable. Furthermore, a design with a fixed period length might be undesirable if the treatment is highly effective, as patients randomized to an initial placebo might experience multiple episodes on placebo. An

alternate design where patients are crossed over to the opposite arm once an event occurs is appealing as it reduces placebo exposure and shortens the overall duration of a clinical trial, while still yielding efficient information from each patient.

We evaluate the power of the various methods under this design. To simulate recurrent events we generate data under a Gamma mixture of Weibulls as before, but where t is now the time since the last event, not the time since the start of the period. Details are given in the appendix, but essentially the first event for each patient is generated in the infinite past, and subsequent events for a patient follow a patient specific Weibull whose hazard potentially changes at the beginning of each treatment period. Thus patients begin Period 1 some time after an event and begin Period 2 immediately after an event or an event-free Period 1. We believe the most plausible Weibull models are constant or increasing hazards, and thus include $k=1$ and $k=3$ in our simulations. We evaluate a crossover trial with periods that last until the first event or three months, whichever comes first, and require patients to have at least four events in the year preceding randomization. This enrichment should ensure a more efficient trial as patients who are unlikely to have events during the trial are excluded. We use a study sample size of 10 patients per sequence.

As before, we evaluate the power when using a Cox test using just the first period as well as Prescott's test, a stratified Cox, FG-W and HR. Each set of parameters is studied with 10,000 replications.

The results are shown in Table 5. Even though the setup was different from the process we consider in Table 3, we again find that the HR and FG-W have very similar power across the parameters of interest, and they perform much better than the other procedures. We conclude that the crossover would have good statistical properties, and that either the FG-W or HR would have excellent power. We prefer the HR because it is very simple to explain and reflects therapies whose primary goal is complete cessation of events, and failing that, delay of events.

7 Example with left censored data

In this section, we apply our method to a crossover trial whose continuous outcome is subject to a limit of detection. This study enrolled twenty two HIV infected patients with detectable viremia into a crossover trial to examine the effect of a statin drug on viral load. Statins disrupt lipid rafts in the blood stream which are important for viral replication. Patients received statin or placebo for 8 weeks followed by a 4 week washout and then received the opposite arm for a final 8 week period. An effective therapy might reduce the amount of virus to below the limit of detection, which is often the threshold point for changing treatment, so that this is qualitatively better than a small but detectable level of virus; thus the HR method has appeal.

Figure 3 displays a scatterplot of Period 1 versus Period 2 viral load values. None of the viral loads is below the limit of detection of 50 copies/ml and the exact versions of HR and FG-W are identical and provide two-sided p-values of .82 (note: the FG-W p-value is .79, when using their normal approximation approach). Viral loads below 1500 are of interest as this is a cutpoint below which patients are thought to have little risk of transmitting the virus

to their partners (Quinn [16]). Thus, to further examine the data and contrast these two procedures, we conducted an additional analysis where $\log_{10}$ viral loads below 3.176 (i.e., $\log_{10}(1500)$) were treated as censored. Under this artificial censoring, the two methods yielded similar results: the FG-W p-value is .83 while the HR p-value is .92.

8 Discussion

In this paper we propose a new method to analyze crossover clinical trials with a survival endpoint. Along the lines of a statistical cure model, our method is developed to better reflect the clinical importance of outcomes by recognizing that avoidance of an event is much better than delaying an event. HR is very simple to compute and is transparent to explain to clinical colleagues. In most of our simulations, the HR approach has statistical performance similar to the FG-W method, however there are situations where the HR has better power including differential censoring by period and cure models. There are some settings, however, where FG-W might shine. First, in the limit of detection problem, where the detection threshold is simply an artifact of the measurement device, the FG-W is preferred. Second, if there is premature dropout (i.e., withdrawal or loss to follow-up prior to the end of the period), the FG-W method can be used without modification. (However, in a typical clinical trial setting, this advantage would only come into play among patients who dropout in the second period, because if they they dropped out in the first period, they would have no second period value and thus would not be included in the FG-W analysis.) Third, the FG-W method yields a treatment effect estimate, along with an associated standard error; estimation has not been worked out for the HR; however, it is possible an approach that would estimate $P(A \text{ is preferred to } B)$ could be developed akin to the method proposed by Follmann [17]. The Prescott procedure also performs surprisingly well in the simulations, except for the Weibull with increasing hazards. Prescott's test breaks down, however, when every patient has events in both periods, so that a strict preference is unassignable.

We did not study Feingold and Gillespie's [3] score transformation test. If this method were modified to use the Wilcoxon two-sample test instead of a t-test, it would share HR's appealing feature of adapting seamlessly to extreme situations. It is likely to perform better than FG-W under differential censoring and cure models. However, as shown by Feingold and Gillespie [3], it has lower power than FG-W in common scenarios, especially as heterogeneity between subjects increases.

The proposed HR method requires that censoring only occurs at the end of each period, that is, that follow-up time is uniform. Of course, in some clinical trials, there will be dropout and a modification to the HR procedure could address this. One method would mimic the approach described independently by Peto [18] and Prentice [19]. Under this approach, patients who were prematurely censored at time $C - \tau$ would randomly draw a failure time from the Kaplan-Meier distribution estimated from both groups and truncated at C . For each imputation, the HR ranks would be calculated and then averaged over a large number of imputations. These average ranks would then be used as scores and a Wilcoxon two-sample test would be performed on the scores.

Crossover clinical trials rarely employ recurrent time to event endpoints; we think trialists should consider such designs more often. Under such a design, our hierarchical ranking approach is a very simple analysis which is intended to reflect what matters most clinically. It is especially appropriate for evaluating a therapy whose primary goal is complete cessation of the event, and when that is not possible, to delay events. Similarly, if a statistical cure model is a plausible framework for therapeutic area, this method should provide good power. The method performs well across a range of statistical scenarios, and also seamlessly adapts to extreme settings. Potential future areas of research include consideration of more complex crossover designs: specified number of multiple events per treatment period, more than two treatment periods, and more than two treatments.

Acknowledgments

We thank our colleague Dr. Anu Ganesan for providing the data on the statin crossover trial. We also thank Dr. Melody Carter and Dr. Dean Metcalfe for consulting us about the anaphylaxis trial, which motivated our interest in this topic, and for helping us describe the anaphylaxis process. We are grateful to our colleagues, Dr. Pamela Shaw and Dr. Michael Fay, for several useful discussions on the topic. Finally, we thank all our reviewers for thoughtful comments which improved our paper.

References

1. Senn S. Cross-over trials in Statistics in Medicine: the first 25 years. *Statistics in Medicine*. 2006; 25:3420–3442.
2. France LA, Lewis JA, Kay R. The analysis of failure time data in crossover studies. *Statistics in Medicine*. 1991; 10:1099–1113. [PubMed: 1876798]
3. Feingold M, Gillespie BW. Cross-over trials with censored data. *Statistics in Medicine*. 1996; 15:953–976. [PubMed: 8783435]
4. Follmann D, Wittes J, Cutler JA. The use of subjective rankings in clinical-trials with an application to cardiovascular-disease. *Statistics in Medicine*. 1992; 11:427–437. [PubMed: 1609177]
5. Brittain E, Palensky J, Blood J, Wittes J. Blinded subjective rankings as a method of assessing treatment effect: A large sample example from the Systolic Hypertension in the Elderly Program (SHEP). *Statistics in Medicine*. 1997; 16:681–693. [PubMed: 9131756]
6. Neaton JD, Wentworth DN, Rhame F, Hogan C, Abrams DI, Deyton L. Considerations in choice of a clinical end-point for aids clinical-trials. *Statistics in Medicine*. 1994; 13:2107–2125. [PubMed: 7846414]
7. Bjorling LE, Hodges JS. Rule-based ranking schemes for antiretroviral trials. *Statistics in Medicine*. 1997; 16:1175–1191. [PubMed: 9179982]
8. Jones, B.; Kenward, MG. *Design and Analysis of Cross-over Trials*. Chapman and Hall; New York: 2003.
9. Senn, S. *Crossover Trials in Clinical Research*. 2. John Wiley and Sons; Chichester: 2002.
10. Prescott RJ. The comparison of success rates in cross-over trials in the presence of an order effect. *Applied Statistics*. 1981; 30:9–15.
11. Gehan EA. A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*. 1965; 52:650–653. [PubMed: 5858975]
12. Mantel N. Ranking procedures for arbitrarily restricted observations. *Biometrics*. 1967; 23:65–78. [PubMed: 6050473]
13. Desu, MM.; Raghavarao, D. *Nonparametric Statistical Methods for Completed and Censored Data*. Chapman and Hall; Boca Raton: 2004.
14. Woolson RF, Lachenbruch PA. Rank tests for censored matched pairs. *Biometrika*. 1980; 67:597–606.
15. Lehmann, EL. *Nonparametrics: Statistical Methods Based on Ranks*. Holden Day; Oakland: 1975.

16. Quinn TC, Wawer MJ, Sewankambo N, Serwack D, Li C, Wabwire-Mangen F, Meehan MO, Lutalo T, Gray RH. Viral load and heterosexual transmission of Human Immunodeficiency Virus Type 1. *New England Journal of Medicine*. 2000; 342:921–929. [PubMed: 10738050]
17. Follmann DA. Regression analysis based on pairwise ordering of patients' clinical histories. *Statistics in Medicine*. 2002; 21:3353–3367. [PubMed: 12407677]
18. Peto R, Peto J. Asymptotically efficient rank invariant procedures. *Journal of Royal Statistical Society, Series A*. 1972; 135(Part 2):185–198.
19. Prentice RL. Linear rank tests with right censored data. *Biometrika*. 1978; 65:167–179.

9 Appendix

Simulating recurrent events for a crossover trial, which must start at a fixed calendar time requires an initial event "sometime" prior to the calendar time enrollment. Let time 0 be the time of randomization/start of Period 1. We adopt the following notation:

- $-\zeta$ denotes the time of the initial recurrent event which occurs in the distant past, prior to Time 0
- τ denotes distance from current event time to Time 0
- $\theta_0 = \lambda$, where λ denotes the patient's underlying frailty which is drawn from a Gamma distribution.
- $\theta_1 = \lambda e^{Z(0)\beta}$
- $\theta_2 = \lambda e^{Z(\tau)\beta}$

We generate the time to the next event, T , as follows, as illustrated by the schematic in Figure 4. We initially set t equal to ζ . We generate a T and see if it falls past zero. If not we set $t_{OLD} = t$ and then obtain $t = t_{OLD} - T$. We repeat this process until t is less than zero, which signifies that an event occurs past baseline. If this event occurs during Period 1 (i.e., $-\tau < t < 0$), then we generate the next event time in Period 2 with an independent Weibull, because the patient begins Period 2 immediately after experiencing an event in Period 1. If the event occurs after Period 2 (i.e., $t < -\tau$), then no additional event times are generated.

We generate T by simulating a uniform (0,1) and then setting T as $F(T)$ where $F()$ is the cdf for T . Recall that $F(t) = 1 - \exp\{-\Lambda(t)\}$ where $\Lambda(t)$ is the integrated hazard. Because the instantaneous hazard changes over the pre-randomization epoch, Period 1, and Period 2, the cumulative hazard has the following piecewise form.

$$\Lambda(t) = \begin{cases} \int_0^t \left(\frac{k}{\theta_0}\right) \left(\frac{s}{\theta_0}\right)^{k-1} ds = \left(\frac{t}{\theta_0}\right)^k, & \text{if } t < \Delta \\ \Lambda(\Delta) + \int_{\Delta}^t \left(\frac{k}{\theta_1}\right) \left(\frac{s}{\theta_1}\right)^{k-1} ds = \left(\frac{\Delta}{\theta_0}\right)^k + \frac{t^k - \Delta^k}{\theta_1^k}, & \text{if } \Delta < t < \Delta + \tau \\ \int_0^{\Delta} \left(\frac{k}{\theta_0}\right) \left(\frac{s}{\theta_0}\right)^{k-1} ds + \int_{\Delta}^{\Delta+\tau} \left(\frac{k}{\theta_1}\right) \left(\frac{s}{\theta_1}\right)^{k-1} ds + \int_{\Delta+\tau}^t \left(\frac{k}{\theta_2}\right) \left(\frac{s}{\theta_2}\right)^{k-1} ds = \left(\frac{\Delta}{\theta_0}\right)^k + \frac{(\Delta+\tau)^k - \Delta^k}{\theta_1^k} + \frac{t^k - (\Delta+\tau)^k}{\theta_2^k}, & \text{if } t > \Delta + \tau. \end{cases}$$

Using the above, we can determine the probability that the next event occurs prior to baseline, during Period 1, or during Period 2. Let u_0 denote the probability that the next event occurs prior to baseline, and u_1 denotes the probability that the next event occurs prior to Period 2. ($u_2 = 1 - u_1 - u_0$ is the probability that the event occurs after Period 1.) These probabilities are:

$$u_0 = 1 - \exp\left(-\left(\frac{\Delta}{\theta_0}\right)^k\right), \text{ and}$$

$$u_1 = 1 - \exp\left[-\left\{\left(\frac{\Delta}{\theta_0}\right)^k + \frac{(\Delta + \tau)^k - \Delta^k}{\theta_1^k}\right\}\right].$$

We note that the cumulative distribution function is $U = 1 - e^{-\Lambda(t)}$. Thus, we generate U from a Uniform (0,1) distribution, and determine T as follows,

$$T = \begin{cases} \theta_0 \{-\ln(1-U)\}^{1/k}, & \text{if } U < u_0 \\ \left[\Delta^k - \theta_1^k \left\{ \ln(1-U) + \left(\frac{\Delta}{\theta_0}\right)^k \right\} \right]^{1/k}, & \text{if } u_0 < U < u_1 \\ \left[(\Delta + \tau)^k - \theta_2^k \left\{ \ln(1-U) + \left(\frac{\Delta}{\theta_0}\right)^k + \left(\frac{(\Delta + \tau)^k - \Delta^k}{\theta_1^k}\right) \right\} \right]^{1/k}, & \text{if } U > u_1. \end{cases}$$

Paired Event Counts

Figure 1.

Counts of patients by occurrence and timing of events in Periods 1 and 2 for the data of Table 1. Counts to the right of the 45° dashed line are for patients who have a preference for Period 1, either because they had no event in Period 1 and an event in Period 2 (N_{01}) or because their time to failure was longer in Period 1 than Period 2 (N_{11}^1). Counts above the diagonal are for patients with this preference for Period 2. Counts in the upper right quadrant are for those patients with no events in either period.

Figure 2. Paired survival times for the 15 patients of Table 1. Thick numbers denote the BA sequence, thin numbers are for the AB sequence. Numbers denote patients' ranks under hierarchical ranking. Lines and shaded areas indicate the region of possible differences for the censored observations.

Figure 3. log₁₀ viral loads of HIV infected individuals in a two period crossover trial at the end of Period 1 (X-axis) and Period 2 (Y-axis). Dots denote patients who had statin in Period 1 followed by placebo. The x's identify the placebo-statin group. Under an effective treatment, the dots would lie above the 45 degree line and the x's below. The length of the line to the 45 degree line reflects the degree of preference.

Figure 4.

Generation of the next event time T for a crossover trial of recurrent events are repeatedly generated until a recurrent event exceeds zero, and thus occurs during (or after) the trial of length 2τ . Once such an event occurs, generation of the next event is trivial: if T falls in Period 1, patients immediately cross over, Period 2 or past it, no further events are generated.

Table 1

Sample data for a two period crossover trial with a time to event outcome. Number of events in periods 1,2 denoted by $k\ell$.

ID	T_1	T_2	$T_2 - T_1$	$k\ell$	Sequence
1	1.00+	0.10	<-.90	01	BA
2	1.00+	0.25	<-.75	01	BA
3	1.00+	0.70	<-.30	01	BA
4	1.00+	0.85	<-.15	01	BA
5	1.00+	0.95	<-.05	01	BA
6	0.95	0.10	-.85	11	BA
7	0.70	0.05	-.65	11	AB
8	0.85	0.55	-.30	11	BA
9	1.00+	1.00+	$(-\infty, +\infty)$	00	BA
10	0.45	0.90	.45	11	AB
11	0.05	0.65	.60	11	AB
12	0.05	0.95	.90	11	BA
13	0.75	1.00+	>.25	10	AB
14	0.55	1.00+	>.45	10	AB
15	0.15	1.00+	>.85	10	AB

Table 2

Data layout for a 2 period crossover trial illustrating Prescott’s test

Sequence	Period Preference		
	Period 1	Neither	Period 2
AB	$N_{01}(AB)$	$N_{00}(AB) + N_{11}(AB)$	$N_{10}(AB)$
BA	$N_{01}(BA)$	$N_{00}(BA) + N_{11}(BA)$	$N_{10}(BA)$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Simulated rejection proportions as a function of the gamma mixing distribution with independent events, $\mu=20$, 50 patients per sequence.

	Test	Design	Exponential $e^{\beta} = 1.83$ $\sigma =$	Weibull k=3 $e^{\beta} = 1.24$ $\sigma =$	Weibull k=.5 $e^{\beta} = 3$ $\sigma =$						
Cox (Period 1 Only)		Parallel	.65	.64	.63	.78	.72	.62	.55	.54	
	Prescott	Crossover	.89	.88	.86	.94	.86	.68	.81	.80	.81
		Crossover	.78	.79	.79	.87	.85	.83	.70	.69	.71
Feingold-Gillespie-W		Crossover	.89	.89	.90	.95	.94	.92	.81	.82	.82
	Heirarchical Rank	Crossover	.89	.89	.90	.97	.95	.91	.81	.81	.82

Table 4
Simulated rejection proportions for special cases; parameter details shown in the text.

Test	Design	Period 2 Censor Factor					Carry-Over	Cure Model
		1	2	5	10	100		
Cox (Period 1 Only)	Parallel	.64	.64	.64	.64	.64	.64	.95
	Crossover	.88	.82	.71	.65	.61	.63	.91
Stratified Cox	Crossover	.79	.70	.49	.36	.20	.53	.62
	Crossover	.89	.79	.53	.32	.02	.63	.63
Feingold-Gillespie-W	Crossover	.89	.82	.72	.67	.61	.65	.82

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Simulated rejection proportions as a function of the gamma mixing distribution with recurrent events, $\mu=.20$, 10 patients per sequence, a minimum of 4 events in year prior to randomization is required.

Table 5

Test	Design	Exponential $\sigma^2=5$		Weibull $k=3$ $\sigma^2=1.65$	
		.01	.05	.01	.05
Cox (Period 1 Only)	Parallel	.63	.66	.68	.60
	Crossover	.84	.85	.77	.66
Stratified Cox	Crossover	.83	.84	.85	.63
	Crossover	.88	.89	.91	.89
Feingold-Gillespie-W	Crossover	.86	.88	.90	.88
	Crossover				.84