STATISTICS IN MEDICINE
Statist. Med. **18**, 1341–1354 (1999)

COMBINING MORTALITY AND LONGITUDINAL MEASURES IN CLINICAL TRIALS

DIANNE M. FINKELSTEIN[1,2]* AND DAVID A. SCHOENFELD[2]

[1] *Biostatistics Department, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.*
[2] *Massachusetts General Hospital, Boston, MA 02114, U.S.A.*

SUMMARY

Clinical trials often assess therapeutic benefit on the basis of an event such as death or the diagnosis of disease. Usually, there are several additional longitudinal measures of clinical status which are collected to be used in the treatment comparison. This paper proposes a simple non-parametric test which combines a time to event measure and a longitudinal measure so that a substantial treatment difference on either of the measures will reject the null hypothesis. The test is applied on AIDS prophylaxis and paediatric trials. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Clinical trials are often designed to compare therapies on the basis of time until the occurrence of an event such as death. Such trials are easy to interpret because the therapies are being compared by their effect on the most important sequelli of the disease. However, clinical trials often consider other measures of treatment effect, and are sometimes designed to use these as the primary measure of treatment efficacy. Sometimes these measures are chosen because it would take too long to accumulate enough deaths to compare the treatments. Alternatively these measures may be chosen because they measure the day to day effect of the disease on the patient. In both cases, data on patient mortality should be combined with these longitudinal measures when treatments are being compared. For example, in a trial of treatments for HIV, therapies are compared on the basis of mortality. However, changes in the CD4 lymphocyte count can be used to measure the effect of treatment, as falling CD4 counts are highly correlated with the declining course of the disease. For trials of therapies to prevent opportunistic infections, the primary measure of efficacy is the occurrence of infections. Although opportunistic infections can be lethal, the treatment comparison is not usually based on mortality, as HIV infected patients die of many often undetermined causes. However, survival should be considered since on the one hand, these therapies could prolong life

* Correspondence to: D.M. Finkelstein, Biostatistics Department, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A. E-mail: dfinkel@sdac.harvard.edu

Contract/grant sponsor: ACTG
Contract/grant number: NO-AI-95030

Contract/grant sponsor: NIH
Contract/grant number: CA-74302

CCC 0277–6715/99/111341–14$17.50

Received January 1998
Accepted August 1998

by preventing a potentially lethal infection, but they could also have unanticipated toxic effects when used over an extended period of time.

Ethical issues also permeate the choice of endpoints. Paediatric AIDS is a disease of long duration leading to progressive declines in growth and neurologic development gradually leading to death. Since it would be unethical to treat patients with progressive disease, survival is not a good measure of treatment benefit because treatments would have been changed long before death occurred. The use of longitudinal measures is also problematic because therapy must be changed when these measures appear to decline, which causes difficulties in making statistical comparisons in longitudinal measures. The usual solution is to compare treatments on the basis of an event which is defined by a complicated combined measure of failure of therapy which includes death, the occurrence of an opportunistic infection, achieving a specified profound degree of neurologic impairment and failure to thrive. When a patient has such an event the patients therapy could be changed. The power of the trial might be improved if the longitudinal measures were combined with the time to event measure.

In this paper, we propose a simple non-parametric statistical test for analysing the impact of treatment which combines a (possibly censored) event with a longitudinal measure of clinical effect. The test is based on a Wilcoxon rank sum test rather than the Cox[1] model. These methods are applied and discussed in the analysis of two AIDS clinical trials of therapies for prevention of opportunistic infections, to compare the treatments on their impact on mortality and the burden (multiplicity) of opportunistic infections. We also present an application of the proposed method for analysis of a paediatric AIDS study comparing two therapies for impact on mortality and growth retardation. The test is introduced in Section 2. Section 3 discusses the adaptation of the test for sequential monitoring. Section 4 illustrates the use of the test for two different AIDS trials. The final section discusses the issues involved in the analysis of a measure that combines a failure (possibly censored) with a longitudinal measure of clinical effect.

2. TESTING FOR A COMBINED LONGITUDINAL MEASURE AND MORTALITY

We wish to have a simple non-parametric statistical test which combines a time to event and a longitudinal measure so that if there is a substantial difference between treatments in either measure the test will reject the null hypothesis. For this, we propose a modification of the generalized Wilcoxon test. The test is based on the principle that each patient in the clinical trial is compared to every other patient in a pairwise manner. For simplicity, we will assume that the failure of interest in the trial is death, and that another longitudinal measure is collected on each patient over time. We define a score, u_{ij}, which is chosen to reflect whether patient i has had the more favourable outcome than patient j. If one of the patients is known to have died before the other, the score is $+1$ if patient i has lived longer than patient j and -1 otherwise. If one cannot tell which patient died first, as in the case where one is censored before the other has died, then we compare the longitudinal measure at the smaller of their follow-up times and assign a $+1$ or a -1 depending on whether or not patient i has a better longitudinal outcome than patient j. If it is not possible to assign either patient as having the better outcome, then the score assigned to the pair is 0.

The proposed test is a score test based on the sum of the scores for the treated group. Suppose there are N subjects in the trial, let $D_i = 1$ for subjects in group 1 and $D_i = 0$ for patients in the other group. Using the u_{ij} for every pair of patients defined above, we assign a score to each

subject, $U_i = \sum_{i \neq j} u_{ij}$. The test is now based on

$$T = \sum_{i=1}^{N} U_i D_i. \tag{1}$$

Without the longitudinal measure this test reduces to Gehan's generalization of the Wilcoxon test. If all patients were followed for the same period of time and there is only a longitudinal measure then this test reduces to the ordinary Wilcoxon test applied to the longitudinal measure at the end of the trial. We will refer to the test as the generalization of the Gehan Wilcoxon test or GGW test. The null hypothesis of this test is that neither survival nor the longitudinal measure is affected by treatment, and the alternative is that at least one and possibly both have improved.

We note that this test is related to the test proposed by Moyè *et al.*[2] The distinction lies in the fact that their test is devised for a study in which there is a quantitative measurement made for each patient who survives the protocol-mandated study duration. Patients are compared on survival if this comparison is possible, and are otherwise compared on the quantitative endpoint if the measurement was made (at the end of the study). By contrast, we are considering a longitudinally collected quantitative measurement, (such as the burden of infections), and for patients who cannot be compared on survival, we will consider the information collected in this measurement for the window of time in which it is available on both (the minimum follow-up of the two). If there is a hierarchy of importance of longitudinal measures, the comparison could be made on a primary measured variable, and if these are comparable, on a secondary variable. For example, for the AIDS prophylaxis study, the comparison would be initially on mortality. If this is not possible (due to censoring or equal death times), then the comparison would be based on the number of infections. If these are comparable, the comparison would be based on a longitudinal measure of quality of life.

Since the use of a stratified test facilitates sequential monitoring and may be useful in other contexts, we describe the distribution of a stratified test. We assume that patients are divided into strata $k = 1, 2, \ldots$, that A_k is the set of indices of the n_k patients in the kth strata and that U_i is calculated within strata (for $i \in A_k$, $U_i = \sum_{j \in A_k} u_{ij}$). Then the test statistic is based on

$$T = \sum_{k} \sum_{i \in A_k} D_i U_i.$$

We find the mean and variance of T conditional on the scores, $U_1 \ldots$, by using the permutation distribution of the treatment assignments within strata, as described in references 3 and 4. We treat $D_1 \ldots$ as random variables and $U_1 \ldots$ as fixed quantities. Let m_k be the total number of patients in the kth stratum who are in the treated group. Then in the kth strata $E(D_i) = \frac{m_k}{n_k}$ and $\mathrm{cov}(D_i, D_j) = \frac{m_k(n_k - m_k)}{(n_k - 1)n_k^2}(\delta_{i,j} n_k - 1)$, where $\delta_{i,j}$ is the Kronecker delta (valued as 1 if $i = j$ and 0 otherwise). Since $\sum_{i \in A_k} U_i = 0$ this implies that the mean of T is zero and its variance is

$$V = \sum_{k} \frac{m_k(n_k - m_k)}{n_k(n_k - 1)} \left(\sum_{i \in A_k} U_i^2 \right).$$

The hypothesis of interest is tested by comparing $\mathscr{Z} = T/V^{1/2}$ to the normal distribution.

In order to calculate the power of this test, we note that the mean of this distribution under the alternative hypothesis requires the calculation of p_k (for subjects in strata k) which is the probability that a patient in group 1 will be doing 'better' than a patient in group 0 at the

minimum of their follow up times. If no ties are possible the mean of the numerator is

$$\sum_k 2m_k(n_k - m_k)\left(p_k - \frac{1}{2}\right)$$

and the mean of denominator is approximately

$$\left\{\sum_k \frac{m_k(n_k - m_k)n_k}{3}\right\}^{1/2}.$$

To calculate p_k one needs to create a model for mortality and for the longitudinal measure. Then p_k can be directly calculated or calculated by simulation. Models used in AIDS clinical trials, with surrogate markers are discussed in references 5 to 10. Models used in cancer are discussed in reference 11. It is clear from the expressions above that the relationship among the longitudinal and mortality measures have an effect on the power of the test through p_k. Therefore, it is important to select the right model and to assess this relationship before utilizing this test for interpretation of a study.

3. ADAPTATION TO SEQUENTIAL MONITORING

The data from clinical trials are reviewed regularly to assess the safety of patients on the study. In order to preserve the overall type I error of the trial, the nominal level for each review (or look) is calculated with consideration of the whole monitoring scheme. For these nominal levels, we find the associated boundary beyond which the value of our test statistic should be rejected and the trial halted. For example, suppose that there are to be three interim reviews. These will each produce a $\mathscr{Z}$-score from the above test. The trial is stopped if

$$\mathscr{Z}_1 > b_1 \text{ or}$$
$$\mathscr{Z}_1 \leqslant b_1 \text{ and } \mathscr{Z}_2 > b_2 \text{ or}$$
$$\mathscr{Z}_1 \leqslant b_1 \text{ and } \mathscr{Z}_2 \leqslant b_2 \text{ and } \mathscr{Z}_3 > b_3.$$

For r looks at the data, we choose boundaries $b_1, b_2, \ldots, b_r$ so that the probability of stopping at some time during the trial is less than 0·05. Thus, associated with each boundary is an increasing sequence $\tau_1, \ldots, \tau_r$ where τ_j is the probability of stopping on or before the jth look. Usually $\tau_r = 0{\cdot}05$. The timing of the looks are based upon the number of events[12] or on calendar time.[13] The choice of τ_j is somewhat arbitrary; the choice suggested in reference 14 is most commonly used.

To determine the boundaries, we must account for correlation in the test statistic from one review analysis to the other. Usually the formula given in reference 15 can be applied to determine boundaries. However, for our test, we do not have independent increments, and thus we will develop an algorithm which simulates the distribution of the sequence of test statistics to calculate the boundaries. The calculation of the sequential form of the test statistic is substantially simplified if the data are stratified by the time the patients enter the study. This is in addition to any stratification based on patient characteristics. Thus, the first strata are patients who enter before the first examination of the data. The second strata includes patients who enter between the first and the second examinations etc. Block sequential randomization stratifies patients by the time they enter a trial, so that this stratification also ensures that the permutation distribution which is used to calculate the distribution of the test statistic is realistic.

If we let i index subject, let k index strata (determined by time of study entry), let A_k be the collection of indices of subjects in the kth strata, let j index the interim review number, let U_{ij} be the value of the score for the ith subject at the jth review, and let

$$T_{kj} = \sum_{i \in A_k} D_i U_{ij} \tag{2}$$

then the test at review j will be based on

$$\mathscr{Z}_j = \frac{\sum_k T_{kj}}{[\sum_k \mathrm{var}(T_{kj})]^{1/2}} \tag{3}$$

where k runs through the indices of strata of patients entered prior to the jth review. The covariance between the test numerator for any two strata k and k' at any two review times j and j', $E[T_{kj}, T_{k'j'}]$ can easily be calculated by noting that the expectation will be 0 for $k \neq k'$. In this case, the covariance is

$$\sum_k \frac{m_k(n_k - m_k)}{n_k(n_k - 1)} \left(\sum_{i \in A_k} U_{ij} U_{ij'} \right). \tag{4}$$

With this simplification, it is not difficult to compute the b_j. The situation for a one-sided boundary is described below, and a two-sided boundary would be computed in a similar way.

We suppose first, for simplicity, that the variance-covariance matrix is known, and we wish to find whether the observed values of $\mathscr{Z}_1 \ldots \mathscr{Z}_r$ crosses a boundary. We begin by generating a sample of $N-1$ vectors from the multivariate normal distribution, which can be calculated using the formulae in (4) above, and which is the asymptotic distribution of our test statistic. We discuss the choice of N later. We include the observed vector to generate a set of N vectors. Then we define an ordering on these N vectors. The ordering is based on using the first component of the N vectors to order the largest $N\tau_1$ values, the next component of the vectors to order the next largest $N(\tau_2 - \tau_1)$ values, and so on. The last $N(\tau_r - \tau_{r-1})$ vectors are ordered based on the last component. The ordering for the remaining $N(1 - \tau_r)$ vectors is not important. The procedure defines an ordering on the N multivariate normal vectors that we generate. The test will reject if the vector of statistics calculated from the data falls in the largest $N\tau_r$ values based on this ordering. The reason the significance level is 0·05 is that if we consider the combined sample consisting of the observed vector of statistics from the data and the other $N-1$ vectors, we have N vectors which are i.i.d. under the null hypothesis. The probability that the observed statistic falls in the largest $\tau_r N$ is the hypergeometric probability of observing one defective in a sample of $\tau_r N$ from a population of N with one defective. This probability is τ_r.

Notice that the size of N, except as defined above, does not play a role. With small N, the proposed test would be a randomized test and some of the uncertainty would come from the generation of the $N-1$ values. Thus, if one had exactly the same data and repeated the procedure several times the results might be different, but the overall rejection probability would equal the nominal probability. If N were too small, then the power could be affected.

If one applies this method to a single 'look', it is easy to calculate by numerical integration the effect of N on power. If a test based on the normal probability table has a power of 0·8, then the proposed method has a power of 0·783 when $N = 100$ and 0·798 when $n = 1000$ at a one-sided significance level of 0·05. Since the generation of large N is relatively easy, one could let $N = 10,000$. For simulations of the method where test statistics needed to be computed multiple times, an N of 500 would probably be adequate.

Since we do not ordinarily have the variance-covariance matrix, we now modify the algorithm as follows. At the first 'look', generate $N-1$ normal random variables with mean 0 and variance equal to the variance of the statistic, which can be calculated using the formula in (4) above. For the observed statistic not to be rejected based on the first component, it would have to be among the smallest $N(1-\tau_1)$ vectors. Thus it would have to less then the smallest value of the largest $N\tau_1$ of the generated values. Let b_1 be this value. To calculate b_2, generate normal random variables for each of the $N(1-\tau_1)$ normal random variables that are $\leqslant b_1$ from the distribution, conditional on the previous $N-1$ values. The conditional distribution is calculated using the variance covariance given in (4). Then calculate b_2 as the smallest value of the largest $N(\tau_2 - \tau_1)$ of these values. This process is repeated for each review. The test using this algorithm will have size very close to τ_r, as long as the sample size of the study is large enough for the asymptotic distribution of the test statistic to be valid and N is large enough so that the $N\tau_j$ are integers.

This method requires that at each look the samples used to find the boundary be saved so that they could be used at the next look. An alternative method which is equivalent to the one above for large N, is to generate $N-1$ multivariate normals at each look using the estimated variance covariance matrix. At the jth look we would go through the procedure based on the ordering of the N vectors, defined above in the description of the method for the case of known variance. There would be a small chance that at the jth look it would appear that we should have stopped the study at a previous look. However, if N is very large this is very unlikely and should be ignored.

We note that when a combined failure and longitudinal outcome is monitored for early closure, the early reviews are driven more by the longitudinal than the failure event because there is more information from these measures. With time, this balance changes. If the longitudinal measure was a surrogate marker, early termination might be called for by marker values alone before any clinical events have occurred. This would lead to problems if the purpose of the trial was to show a clinical benefit. Owing to this, caution must be exercised in utilizing the combined measure. It is most appropriate to use a combined measure in the case that the treatment comparison is based on an event which is defined by a longitudinal measure crossing a threshold and the longitudinal measure is monotone with time among untreated patients. For example, for a paediatric study where a crossing growth failure threshold defines failure, the early indication of steeply falling growth parameters would indicate a high likelihood of eventual failure, and thus it would be reasonable to utilize the longitudinal information from each patient in deciding early termination of a trial. The decision of monitoring on the combined measure would be made at the design stage, and requires confidence that such results should drive early termination.

4. EXAMPLES

4.1. Analysis of paediatric AIDS study

ACTG 152 compared AZT versus ddI versus combination therapy in children with AIDS.[16] The study accrued 831 patients and was conducted between 1991 and 1995. The mean follow-up for patients on the trial was 32 months. Children had various parameters such as weight, head circumference and height measured periodically. A failure was said to occur (and treatment was stopped) if the patient died, developed an opportunistic infection, had CNS deterioration or growth failure. For younger children, head circumference is a particularly sensitive measure of failure to thrive. We will combine measurements of head circumference with the study defined failure to produce a combined efficacy measure.

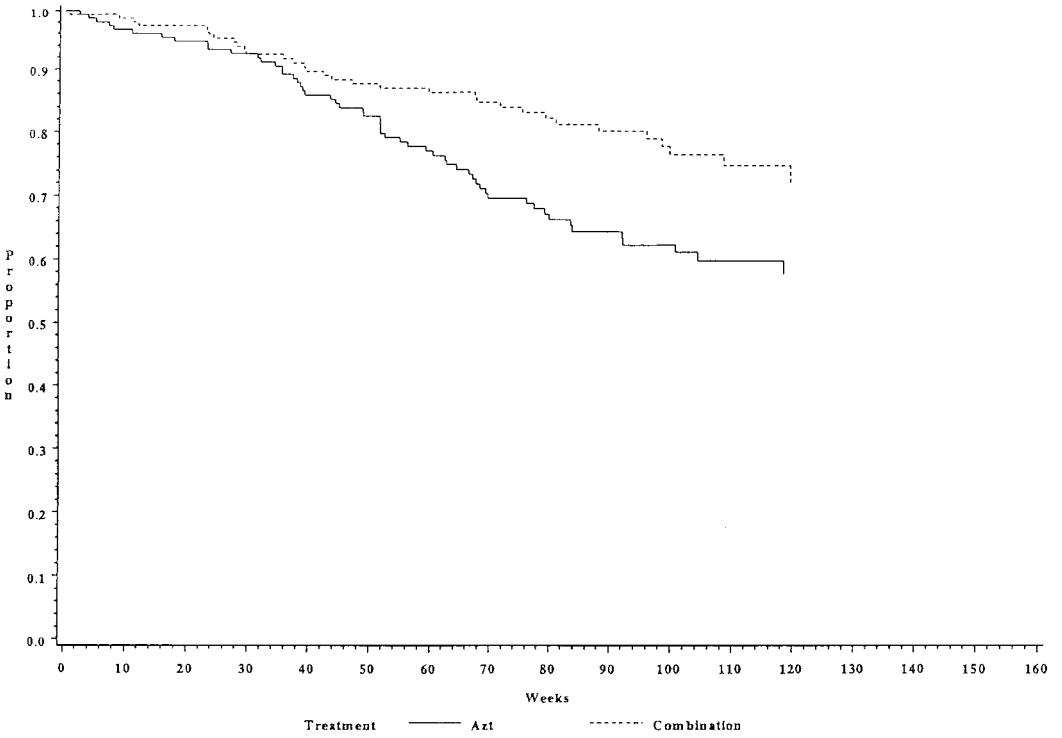

Figure 1. Treatment comparison of progression from ACTG 152

To do this, we first convert head circumference to a Z-score based on the normal head circumference for children of the same age, which was available by week. Then the difference of this Z-score from the baseline Z-score is computed. To simplify the calculation, we average these scores every eight weeks and use the previous average value at each follow-up time. For instance if the minimum follow-up time for two patients is 12 weeks, then we use their eight week value, which would be the average of any measurements taken from week 4 to week 12. The calculation of the statistic is performed in SAS by creating a vector of observation values for each patient as well as their failure and/or follow-up time and an indicator of their vital status.

We concentrate on the difference between AZT and the combination of AZT and ddI. Among younger children there was a significant decrease in the hazard for time to a primary failure ($Z = 2{\cdot}75$, $p = 0{\cdot}006$, logrank test) for patients on the combination therapy. The two failure time curves are shown in Figure 1. There was also an apparent difference in the longitudinally measured head circumference Z-scores shown in Figure 2, but this is difficult to test due to the different length of follow-up for each patient. The combined measure of failure showed a highly significant difference between the groups ($Z = 3{\cdot}54$, $p = 0{\cdot}0004$, GGW test). This test has a more significant p-value, resulting from the fact that the GGW test had a larger effect size than the logrank test for detecting treatment differences. We note that in this trial the conclusions of treatment effect were not altered by using the GGW test, as the logrank test was already significant. However, clearly the combined measure of failure analysis had a larger effect size, and in a study with

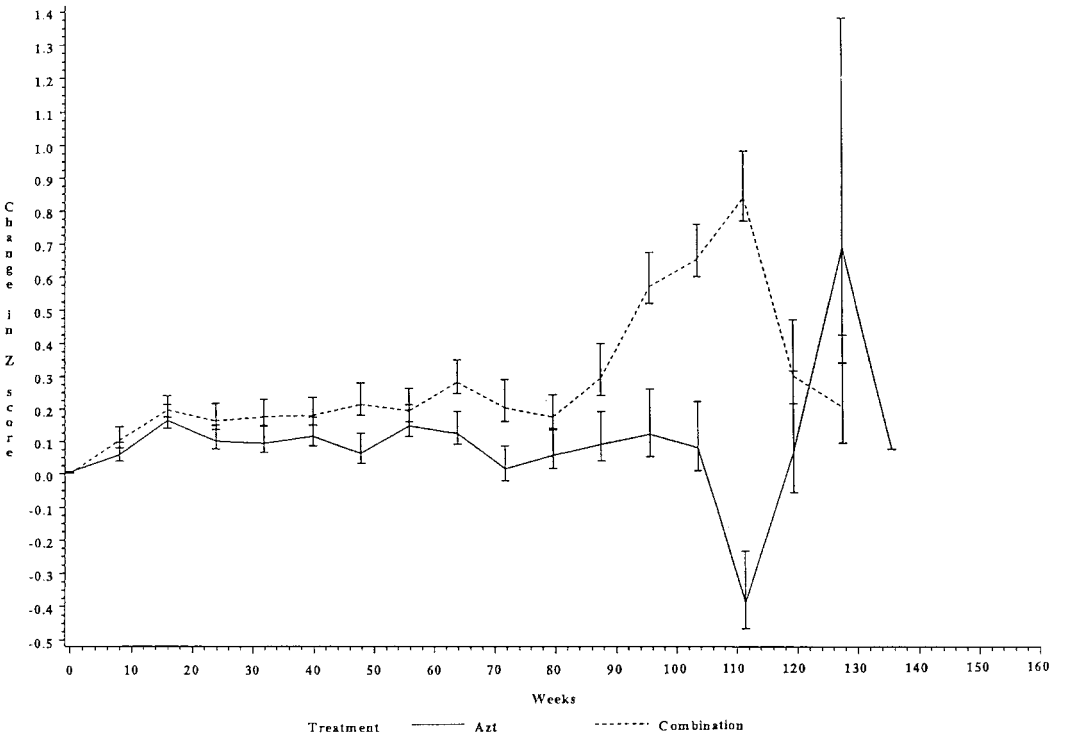

Figure 2. Treatment comparison of failure to thrive (changes in head circumference) from ACTG 152

marginally significant mortality differences, this test could change the conclusions of the study. One should note that in this study each of the measures that made up the combined endpoint nominally favoured the combination therapy (data not shown).

4.2. Analysis of multiple PCP in ACTG 081

ACTG 081 was a phase III comparison of Trimethoprim-Sulfamethoxazole (TS), Dapsone, and aerosolized Pentamidine for prevention of primary *Pneumocystis carinii* pneumonia (PCP). The protocol specified that patients were to remain on treatment even if they experienced an infection. There was no expectation that there would be a survival advantage to any of these therapies. In the absence of mortality differences, the primary measure of treatment effect in the study was to be the occurrence of a PCP. This discussion will focus on the comparison of TS, the standard systemic therapy, and AP, a localized therapy.

ACTG 081 accrued 280 patients per arm beginning in 1989, and closed after a median follow-up of approximately 3 years. The entry criteria was CD4 counts below 200, and 185 patients in the TS or AP arms entered with CD4 counts below 100. As reported in Bozette *et al.*,[17] there was no significant treatment difference when all patients were included in the analysis. However, in the subgroup of patients with CD4 counts below 100, TS was found to be more effective than AP in preventing a first PCP ($p = 0{\cdot}04$). Patients on TS experienced a longer survival than those on AP, but the difference was not statistically significant. The publication from this trial reported

Table I. Recurrences of PCP in patients with CD4 < 100 on ACTG 081

Treatment	Number of PCP			
	0	1	2	3
AP	64	19	4	2
TS	80	13	3	0

Table II. Treatment comparisons based on GGW test

Analysis	Z	p-value
First PCP	1·9241	0·0528
First PCP/death	2·1893	0·0286
Recurrent PCP	1·9855	0·0470
Recurrent PCP/death	2·2083	0·0272

the comparison of the treatments on the basis of survival and PCP-free survival. The primary outcome was the first occurrence of PCP. However, as shown in Table I, there were several PCP recurrences experienced by some patients in this trial.

We applied the GGW test to these data. For treatment comparison on the time to first PCP, as Table II indicates, the TS therapy was more effective in preventing the first PCP infection. Note that the p-value for this comparison based on the GGW test was not identical to that from the logrank test, but the differences were not very large. The GGW methodology was applied next to a hierarchical comparison based first on mortality, then on recurrences of PCP. Therefore, for two patients whose mortality comparison was not possible due to censoring, the comparison was made on the basis of the number of PCP experienced by the minimum of their follow-up time. As Table II indicates, inclusion of mortality resulted in a more significant p-value, as did inclusion of number of recurrences. However, the hypothesis which compared the treatments on the basis of all information on recurrences and survival provided the most powerful comparison.

4.3. Analysis of multiple events in ACTG 981

ACTG 981 was a phase III comparison of fluconazole versus clotrimazole troches for prevention of invasive fungal infections in patients with HIV. Fluconazole is a systemic treatment, while clotrimazole troches primarily provide oral protection against candidiasis (thrush) infections. The protocol specified that patients were to remain on treatment even if they experienced an infection. There was no expectation that there would be a survival advantage to either of these therapies. In the absence of mortality differences, the primary measure of treatment effect in the study was to be the occurrence of a *serious* infection including the rare, and sometimes fatal, systemic fungal infections (such as histoplasmosis and cryptococcosis) as well as oesophageal candidiasis.

The study accrued 428 patients between September 1989 and September 1992 and closed to follow-up in June 1993 after a median follow-up of 34·7 months. As reported in Powderly *et al.*,[18] the patients randomized to fluconazole experienced a significant delay in the time to first serious fungal infections (systemic infections or oesophageal candidiasis), compared to clotrimazole ($p < 0{\cdot}0001$, see Figure 3). The analysis of survival by treatment indicated that patients

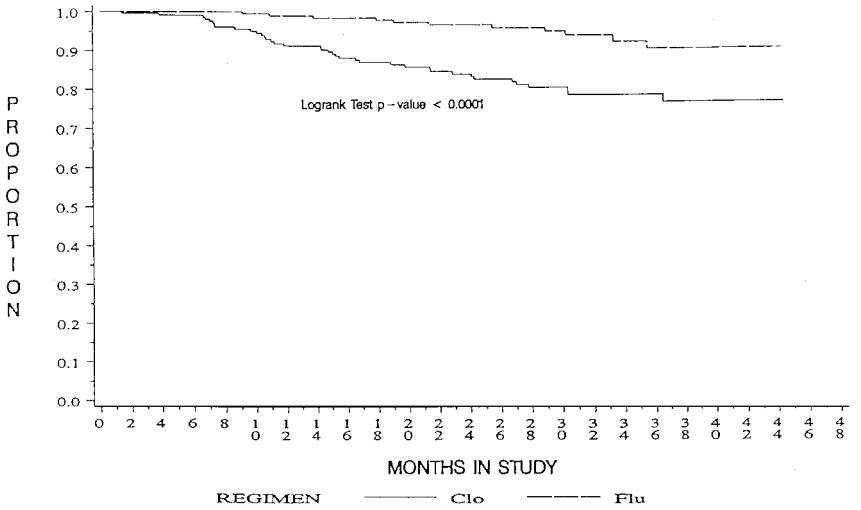

Figure 3. Treatment comparison of occurrences of serious infection from AIDS prophylaxis study ACTG 981

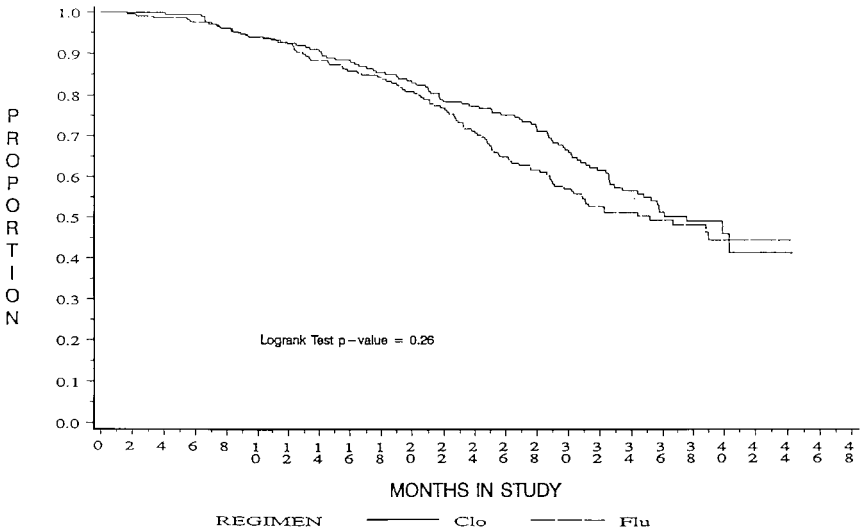

Figure 4. Treatment comparison of mortality from AIDS prophylaxis study ACTG 981

assigned to clotrimazole had a slightly lower mortality risk than those assigned to fluconazole ($p = 0{\cdot}26$), see Figure 4. In light of the fact that mortality risk was higher in the arm that had shown the greatest benefit for prevention of fungal infection (fluconazole), the single most important analysis of the trial was the treatment comparison of infection-free survival time, which did not show either treatment to be superior (see Figure 5).

The report of this study[18] focused on treatment comparisons with respect to the hazard for the event of fungal infection or the combined event of fungal infection or death. However, in some

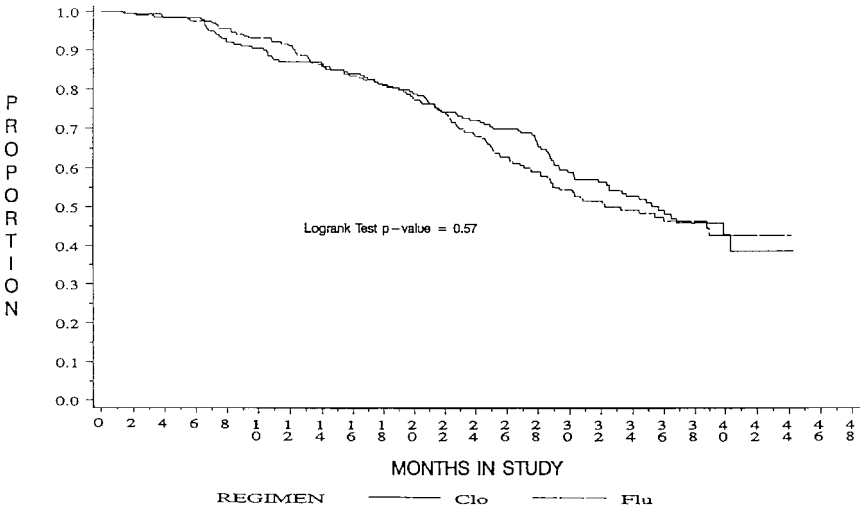

Figure 5. Treatment comparison on basis of infection-free survival time from ACTG 981

sense, the true *burden* of infection could be viewed as the multiple occurrences of these events. The analysis of interest was to compare the treatments on the basis of the occurrence of various types of infections (for example fungemia and oesophageal candidiasis).

We first apply the GGW test to compare treatments on the basis of infections alone, treating death as non-informative censoring. The results indicate that there is a highly significant benefit of fluconazole for prevention of infections ($p < 0{\cdot}0001$). We next apply the GGW test, for a hierarchical comparison. For this, we compare first on mortality, and if this comparison does not distinguish them, we wish to descend on the hierarchy to compare on the basis of the burden of infection. The results of this test indicate that the treatments are not significantly different ($p = 0{\cdot}2439$). Thus, the results change dramatically with the inclusion of information on mortality. We note that this is similar to what is found by applying the multivariate test proposed in reference 19 using reference 20 to the event of mortality or infection versus infection alone (see reference 21).

Finally, we wanted to include quality of life in the hierarchy. Thus, we wanted to reanalyse the data so that for subjects who could not be distinguished in terms of earlier death or more recurrences of infections, we compare on the basis of quality of life. Quality of life was assessed using the PARSE instrument (Rand Corporation),[22] and was evaluated every six months. The instrument was a patient's self-assessment of questions relating to functional, social and psychological status. For this analysis, we used the per cent change in quality of life from baseline, which is shown in Figure 6. It is hard to compare the treatments on the basis of this graph, as patients had varied length of follow-up. We first consider the impact of quality of life in addition to each of the other dimensions of death and recurrences. Using the GGW test on quality of life alone, we see that there is no statistically significant treatment difference ($p = 0{\cdot}9620$). When death and quality of life are considered only, the treatments are not significantly different ($p = 0{\cdot}2171$, GGW test), but when quality of life and infection only are considered, the treatments are significantly different ($p < 0{\cdot}0001$). When all three measures were simultaneously considered (in the order of death, then multiple infections, and last, quality of life), there was no significant treatment difference

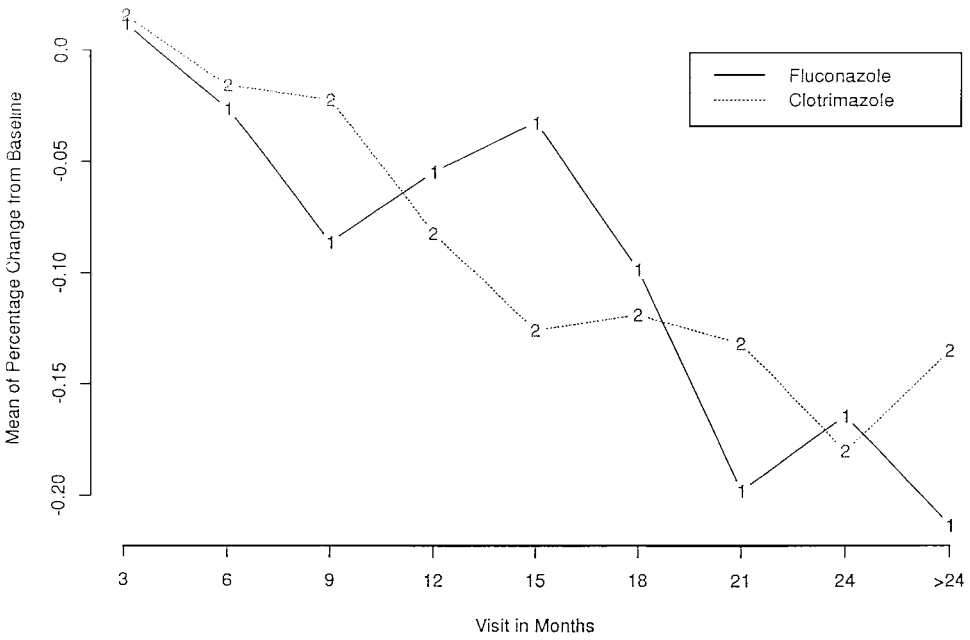

Figure 6. Treatment comparison of quality of life from AIDS prophylaxis study ACTG 981

($p = 0{\cdot}4367$). From this progression of inclusion of each measure of efficacy in this analysis, it is clear that the treatments only really differ on the prevention of infection, which does not translate into a survival benefit. Since quality of life is also not impacted, the conclusion that is drawn is that the more expensive treatment is not providing sufficient benefit in this population of patients.

5. DISCUSSION

Any test that combines different measures of treatment efficacy will assign a relative weight to each measure. Such an assignment can occur explicitly with the use of a utility function such as TWiST[23] or Q-TWiST,[24] or implicitly as part of the estimation of a common parameter of a joint model for both measures.[25] Here we assign weights hierarchically based on an ordering of the relative severity of the possible outcomes.

If our method is to be used, its statistical properties should be examined under different, biologically plausible, alternative distributions by simulation to ensure that a treatment that makes survival worse but improves some less important longitudinal marker will not appear effective. Such a simulation would also be necessary for a combined endpoint based on a utility function because the fact that different measures change at different times in a sequentially analysed clinical trial may effect the relative weighting of the measures as much as the weight function that is chosen. For instance, suppose that TWiST is compared between treatments in a short clinical trial. Then treatments with less early toxicity or a quicker symptomatic improvement will look better, while in a trial with a greater duration of follow-up, the proportion who are cured will dominate any comparison. An analysis of multiple endpoints using a model cannot be misleading in this manner as long as the model is correct. This is not very reassuring considering that most models are

chosen for their mathematical convenience rather than their biological plausibility. A simulation under conditions where the model does not hold serves the same function as the simulations we suggest when our method is used to combine outcomes.

Combining measures that have the same clinical meaning is always appropriate. For instance, in the paediatric trial, the failure which was used as the primary measure of efficacy was itself a combination of several longitudinal markers. There was an ethical reason for defining such a failure rather than just comparing the longitudinal markers. When failure occurred, the child's physician could change medications without violating the protocol. In this situation combining the actual longitudinal measure with the failure could be expected to increase the power of the study without introducing the possibility of misleading results.

Combining measures that have a different clinical meaning is more problematic. For instance, in ACTG 081 and 981 the longitudinal measure is number of infections, which, although clinically important, are not comparable to death. If the true treatment effect is to improve a longitudinal measure but make survival worse, then a combined efficacy measure may lead to the conclusion that the treatment with worse survival is superior. The fact that mortality is included in the efficacy measure makes this less likely than an analysis that ignores mortality, but it can still happen, especially if the effect on the longitudinal measure is strong and occurs early in the trial. This possibility must be ruled out based on prior knowledge in any trial that uses any other measure of efficacy than the mortality rate.

In ACTG 081, where the treatments are intended to prevent PCP, a potentially life-threatening infection, the combination of mortality and number of infections can potentially lead to increased statistical power. An ineffective prophylaxis might lead to lethal infections so both mortality and number of infections are likely to differ in the same direction between treatments and each part of the combined measure will add power to the treatment comparison.

If the true treatment effect is to change a longitudinal measure but it has no effect on mortality, then a combined measure may have less power than a measure that censors the data at death. ACTG 981 may be an example of this phenomenon. Serious fungal infections were rare and contributed to the death of very few patients. On the other hand, mortality itself was common with deaths from a myriad of other opportunistic illnesses. Thus the mortality difference from a difference in the rate of fungal infections would be small compared to the noise from all the other sources of mortality that are not addressed by the study. One might argue that the fact that fungal infections have such a minor affect on mortality argues against their use as an anti-fungal prophylaxis in the first place unless one is absolutely sure that they are benign medications.

ACKNOWLEDGEMENTS

We wish to thank the patients and investigators who were involved in the ACTG 981 study. Support for this research came from NIH grant CA-74302 and the ACTG, under contract NO-AI-95030.

REFERENCES

1. Cox, D. 'Regression models and life-tables (with Discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
2. Moyé, L. A., Davis, B. R. and Hawkins, C. M. 'Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint', *Statistics in Medicine*, **11**, 1705–1717 (1992).
3. Cox, D. and Hinkley, D. *Theoretical Statistics*, Chapman and Hall, London, 1974, p. 185.

4. Mantel, N. 'Ranking procedures for arbitrarily restricted observation', *Biometrics*, **23**, 65–78 (1967).
5. Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. 'Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients With Aids', *Journal of the American Statistical Association*, **90**, 27–37 (1995).
6. Lange, N., Carlin, B. P. and Gelfand, A. E. 'Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (disc: P626-632)', *Journal of the American Statistical Association*, **87**, 615–626 (1992).
7. Taylor, J. M. G. and Segal, M. R. 'Comment on "Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers"', *Journal of the American Statistical Association*, **87**, 628–631 (1992).
8. Zeger, S. L. and Diggle, P. J. 'Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters', *Biometrics*, **50**, 689–699 (1994).
9. Lin, D. Y., Fischl, M. A. and Schoenfeld, D. A. 'Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in Human Immunodeficiency Virus clinical trials', *Statistics in Medicine*, **12**, 835–842 (1993).
10. Satten, G. A. and Longini, Ira M., Jr. 'Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease (disc: P295-309)', *Applied Statistics*, **45**, 275–295 (1996).
11. Lagakos, S. W., Sommer, C. J. and Zelen, M. 'Semi-markov models for partially censored data', *Biometrika*, **65**, 311–318 (1978).
12. Lan, K. K. G. and DeMets, D. L. 'Changing frequency of interim analysis in sequential monitoring', *Biometrics*, **45**, 1017–1020 (1989).
13. Lan, K. K. G. and DeMets, D. L. 'Group sequential procedures: Calendar versus information time', *Statistics in Medicine*, **8**, 1191–1198 (1989).
14. O'Brien, P. C. and Fleming, T. R. 'A multiple testing procedure for clinical trials', *Biometrics*, **35**, 549–556 (1979).
15. Armitage, P., McPherson, C. K. and Rowe, B. C. 'Repeated significance tests on accumulating data', *Journal of the Royal Statistical Society*, *Series A*, **132**, 235–244 (1969).
16. Englund, J. A., Baker, C., Raskino, C., McKinney, R. E., Lifschitz, M. H., Petrie, B., Fowler, M., Connor, J. D., Mendez, H., O'Donnell, K. and Wara, D. W. 'Clinical and laboratory characteristics of a large cohort of symptomatic, human immunodeficiency virus-infected infants and children. AIDS clinical trials group protocol 152 study team', *Paediatric Infectious Disease Journal*, **15**, 1025–1036 (1996).
17. Bozzette, S., Finkelstein, D., Spector, S., Frame, P., Powderly, W., He, W., Phillips, L. D. C., van der Horst, C. and Feinberg, J. 'A randomized trial of three anti-Pneumocystis agents in patients with advanced HIV infection', *New England Journal of Medicine*, **332**, 693–699 (1995).
18. Powderly, W., Finkelstein, D., Feinberg, J., Frame, P., He., W., van der Horst, C., Koletar, S., Eyster, E., Carey, J., Waskin, H., Hooton, T., Hylsop, N., Spector, S. and Bozzette, S. 'A randomized trial comparing fluconazole with clotrimazole troches for the prevention of fungal infections in patients with advanced HIV infection', *New England Journal of Medicine*, **332**, 700–705 (1995).
19. Wei, L. J., Lin, D. Y. and Weissfeld, L. 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association*, **84**, 1065–1073 (1989).
20. Lin, D. Y. 'MULCOX A computer program for the Cox regression analysis of multiple failure time variables', *Computer Methods and Programs in Biomedicine*, **32**, 125–135 (1990).
21. Finkelstein, D., Schoenfeld, D. and Stamenovic, E. 'Analysis of multiple failure time data from an AIDS clinical trial', *Statistics in Medicine*, **16**, 951–961 (1997).
22. Bozzette, S., Kanouse, D., Berry, S. and Duan, N. 'Health status and function with zidovudine or zalcitabine as initial therapy for AIDS: A randomized controlled trial', *Journal of the American Medical Association*, **273**, 295–301 (1995).
23. Gelber, R. D., Gelman, R. S. and Goldhirsch, A. 'A quality-of-life-oriented endpoint for comparing therapies', *Biometrics*, **45**, 781–795 (1989).
24. Gelber, R. D., Cole, B. F., Gelber, S. and Goldhirsch, A. 'Comparing treatments using quality-adjusted survival: The Q-TWiST Method', *American Statistician*, **49**, 161–169 (1995).
25. Finkelstein, D. M. and Schoenfeld, D. A. 'Analysis of multiple tumor data from a rodent carcinogenicity experiment', *Biometrics*, **45**, 219–230 (1989).

