STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:705–731

Efficiency comparisons of rank and permutation tests based on summary statistics computed from repeated measures data

Janice M. Weinberg[1,*,†] and Stephen W. Lagakos[2]

[1] *Department of Epidemiology and Biostatistics, Boston University School of Public Health, 715 Albany Street, Boston, Massachusetts 02118, U.S.A.*
[2] *Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*

SUMMARY

A popular method of using repeated measures data to compare treatment groups in a clinical trial is to summarize each individual's outcomes with a scalar summary statistic, and then to perform a two-group comparison of the resulting statistics using a rank or permutation test. Many different types of summary statistics are used in practice, including discrete and continuous functions of the underlying repeated measures data. When the repeated measures processes of the comparison groups differ by a location shift at each time point, the asymptotic relative efficiency of (continuous) summary statistics that are linear functions of the repeated measures has been determined and used to compare tests in this class. However, little is known about the non-null behaviour of discrete summary statistics, about continuous summary statistics when the groups differ in more complex ways than location shifts or where the summary statistics are not linear functions of the repeated measures. Indeed, even simple distributional structures on the repeated measures variables can lead to complex differences between the distribution of common summary statistics of the comparison groups. The presence of left censoring of the repeated measures, which can arise when these are laboratory markers with lower limits of detection, further complicates the distribution of, and hence the ability to compare, summary statistics. This paper uses recent theoretical results for the non-null behaviour of rank and permutation tests to examine the asymptotic relative efficiencies of several popular summary statistics, both discrete and continuous, under a variety of common settings. We assume a flexible linear growth curve model to describe the repeated measures responses and focus on the types of settings that commonly arise in HIV/AIDS and other diseases. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Consider a randomized clinical trial in which the individuals from two comparison groups have an outcome measured several times over the course of a study. For example, individuals enrolled in an HIV/AIDS clinical trial are commonly evaluated for HIV-1 viral load levels

[*] Correspondence to: Janice M. Weinberg, Department of Epidemiology and Biostatistics, Boston University School of Public Health, 715 Albany Street, Boston, Massachusetts 02118, U.S.A.
[†] E-mail: janicew@bu.edu

Contract/grant sponsor: NIH; contract/grant number: AI24643

Received November 1999
Accepted March 2000

at baseline and at regular intervals thereafter. A popular method of analysis in this setting is to summarize each individual's outcomes with a scalar summary statistic, and then perform a two-group comparison of the resulting statistics using a linear permutation or rank test. Although this approach is used extensively in a variety of disease settings, little is known about the relative performance of tests based on different summary statistics.

The summary statistic approach [1, 2] has many attractive features, especially descriptive simplicity, validity under the null hypothesis whenever missing data occur non-informatively, and, in some instances, when data is missing informatively, by stratifying by missingness pattern [3, 4]. Dawson and Lagakos [5] compare continuous summary statistics that are linear combinations of a repeated measures process, under the assumption that the repeated measures processes for the two groups differ by a location shift at each time point. For this setting they show that the asymptotic relative efficiency (ARE) comparing the two summary statistics of this type, when using the same linear rank test, is a simple function of the vector of mean outcomes in each group and a common covariance matrix. However, except for very specialized situations such as this, even simple summary statistics will generally lead to distributions which differ in more complex ways than location or scale shifts. In addition, the repeated measure response can be left censored when this response is evaluated by a laboratory assay with a lower limit of detection. This feature, which now arises commonly in HIV/AIDS studies as a result of highly active antiretroviral therapies, complicates the distributions of summary statistics and hence their comparison. Recently, Weinberg and Lagakos [6, 7] derived the asymptotic distribution of linear rank and linear permutation tests under general contiguous alternatives to the null hypothesis of group equality. These results can be used to assess the large-sample behaviour of a permutation or rank test based on a summary statistic computed from a repeated measures process, under general assumptions about how the repeated measures processes differ between the comparison groups and how the summary statistic depends on the repeated measures process. These results provide the theoretical basis for the comparisons provided in the present paper, which are intended to provide a basis for the selection and use of summary statistics when comparing treatment groups based on a repeated measures response process.

In Section 2 we introduce some motivating examples from two recently completed HIV/AIDS clinical trials. In Section 3 we present the notation, assumptions and theoretical results used in subsequent sections. In Section 4 we describe a flexible linear growth curve model for describing the behaviour of the repeated measures process, and in Section 5 we describe the summary statistics under consideration. In Section 6 we compare the relative efficiencies of several discrete and continuous summary statistics computed from repeated measures, using popular linear permutation and rank tests, under a variety of treatment difference scenarios. The effect of values dropping below a lower limit of quantification, a form of left censoring, is also addressed. In Section 7 we analyse the data for the examples in Section 2 using the summary statistics described in Section 5 and compare these results to the asymptotic relative efficiency (ARE) comparison of summary statistics discussed in Section 6.

2. EXAMPLES

Consider an HIV/AIDS clinical trial where HIV-1 viral load level is the primary outcome measured repeatedly over time. As mentioned previously, a popular method of analysis in this setting is to summarize each individual's outcomes with a scalar summary statistic, and then

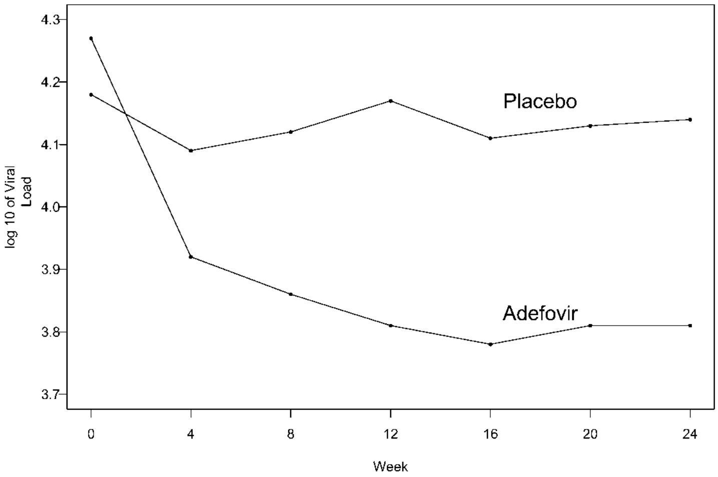

Figure 1. Mean viral load by treatment group: Gilead Science Study 408, patients not taking a protease inhibitor at time of randomization.

perform a two-group comparison of the resulting statistics using a linear permutation or rank test. We present three examples from two recently completed HIV/AIDS clinical trials where the treatment effects over time take on very different forms.

We first consider a trial by Kahn *et al.* in which HIV-infected persons receiving standard-of-care protease inhibitor and nucleoside analogue reverse transcriptase inhibitor drugs were randomized to receive an experimental nucleotide inhibitor – Adefovir dipivoxil – or a placebo for 24 weeks [8]. We restrict our attention to those patients with baseline viral load above 5000 copies/mL. Values falling below the limit of quantification of the assay, 500 viral copies/mL, were replaced by this limit. We first examine the subgroup of 199 patients that were not taking protease inhibitors at the time of randomization, of which 110 were randomized to placebo and 89 to Adefovir. At the end of the study, 14.6 per cent of patients in the Adefovir group and 7.3 per cent of patients in the placebo group are below the lower limit of quantification. Mean $\log_{10}$ viral copies over time, by treatment group, are presented in Figure 1. Here the treatment effect occurs early and is maintained for the 24-week study period.

Next consider the subgroup of 143 patients taking a protease inhibitor at the time of randomization, of whom 71 were randomized to placebo and 72 to Adefovir. At the end of the study, 15.3 per cent of patients in the Adefovir group and no patients in the placebo group are below the lower limit of quantification (see Figure 2). In contrast to the results seen in Figure 1, the treatment effect occurs primarily towards the end of the study, with the largest difference seen at week 24.

As another example, Study 241 of the AIDS Clinical Trials Group was a randomized trial that evaluated the effect of adding the drug nevirapine to a common two-drug regimen consisting of the drugs AZT and ddC on short- and long-term viral load levels in patients infected with HIV [9]. Other studies also have corroborated the antiviral effect of nevirapine. There are a total of 203 patients, with 103 randomized to triple therapy (AZT + ddC + nevirapine) and 100 randomized to double therapy (AZT + ddC). Values falling below the limit of quantification of the assay, 200 viral copies/mL, were replaced by this limit. At the

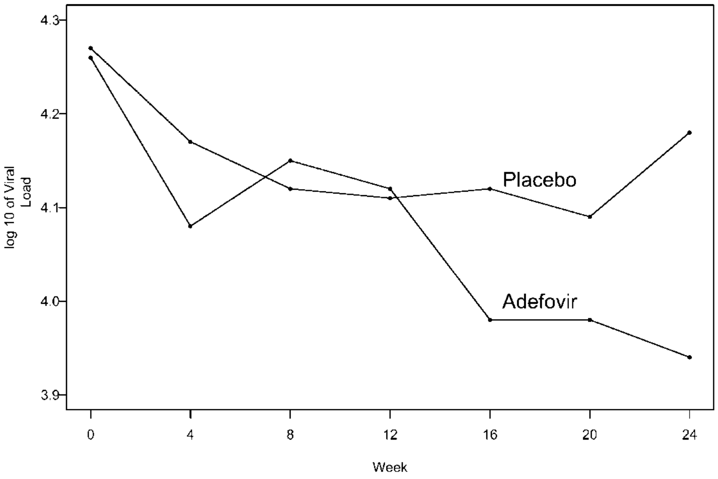

Figure 2. Mean viral load by treatment group: Gilead Science Study 408, patients taking a protease inhibitor at time of randomization.

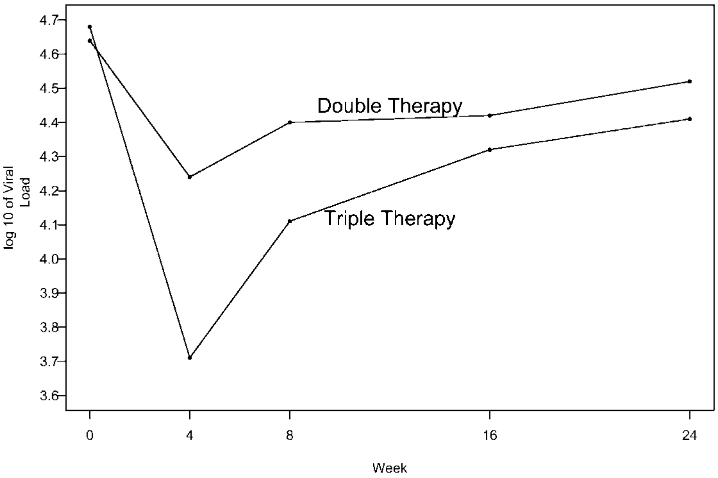

Figure 3. Mean viral load by treatment group: ACTG 241.

end of the study, 10.7 per cent of patients taking triple therapy and 7.0 per cent of patients taking double therapy were below the lower limit of quantification. As seen in Figure 3, there is a short-term treatment effect by week 4, which diminishes by the end of the study period, presumably due to the development of nevirapine resistance.

For each of these settings it is not clear which test based on a summary statistic will best detect treatment differences or how different summary statistics will perform. We return to these examples in Section 7 where we analyse the data using the summary statistics described in Section 5 and compare these results to the asymptotic relative efficiency (ARE) comparison of summary statistics discussed in Section 6.

3. NOTATION, ASSUMPTIONS AND TEST STATISTICS

Let X_i denote the summary statistic for the ith of $m+n$ individuals, and assume that $X_1,\ldots,X_{m+n}$ are independent, with distribution function $F(x|\boldsymbol{\theta}_0)$ for $i=1,\ldots,m$ and $F(x|\boldsymbol{\theta}_1)$ for $i=m+1,\ldots,m+n$, where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are both $P\times 1$ vectors of parameters. We refer to the two sets of observations as groups 0 and 1 and are interested in testing the null hypothesis $H_0:\boldsymbol{\theta}_0=\boldsymbol{\theta}_1$. It is assumed that the groups will be compared using a linear permutation test when X_i is discrete and either a linear rank or permutation test when X_i is continuous. The test statistic corresponding to a linear rank test is of the form

$$T_N^R=\frac{1}{m}\sum_{i=1}^{m}a_N(R_{Ni})$$

where $N=m+n$ is the total sample size, R_{Ni} is the rank of X_i among $X_1,\ldots,X_N$ and $a_N(1),\ldots,a_N(N)$ are values of a score function $a_N(.)$. Here and elsewhere we use (.) to refer to an entire function. For example, use of $a_N(R)=R$ corresponds to the Wilcoxon two-sample test (see Randles and Wolfe [10]).

A linear permutation test is of the form

$$T_N^P=\frac{1}{m}\sum_{i=1}^{m}a(X_i)$$

where $a(.)$ is some known function or score. In practice, the most commonly-used transformation is the identity function $a(X)=X$.

We consider the asymptotic behaviour of the standardized test statistics under a general sequence of contiguous alternatives $H_1:\boldsymbol{\theta}_1=\boldsymbol{\theta}_1(N)=\boldsymbol{\theta}_0+\boldsymbol{\delta}/\sqrt{N}$ where $\boldsymbol{\theta}_0$ and $\boldsymbol{\delta}$ are fixed. For linear rank tests, $F(x|\boldsymbol{\theta}_1)$ and $F(x|\boldsymbol{\theta}_0)$ are restricted to be absolutely continuous distribution functions which are equal when $\boldsymbol{\theta}_1=\boldsymbol{\theta}_0$. If μ_{N0} and σ_{N0}^2 denote the mean and variance of T_N^R under H_0, then Weinberg and Lagakos [6] show that under H_1 and regularity conditions

$$\frac{T_N^R-\mu_{N0}}{\sigma_{N0}}\xrightarrow{\mathscr{L}}N(\xi,1)\quad\text{as } N\to\infty$$

where

$$\xi=\frac{\sqrt{\{\lambda(1-\lambda)\}}\sum_{p=1}^{P}\delta_p\int_0^1\varphi'(u)K_p\{F^{-1}(u)\}\,\mathrm{d}u}{\sqrt{[\int_0^1\{\varphi(u)-\bar{\varphi}\}^2\,\mathrm{d}u]}}\tag{1}$$

$$F(.)=F(x|\boldsymbol{\theta}_0),\quad K_p(x)=\frac{\partial F(x|\boldsymbol{\theta}_0)}{\partial\theta_{0p}},\quad p=1,2,\ldots,P,\quad \lambda=\lim_{N\to\infty}\frac{m}{N}\in(0,1)$$

where $\lim_{N\to\infty}a_N(1+[uN])=\varphi(u)$, $0<u<1$ and $\bar{\varphi}$ is the expected value of $\varphi(u)$.

For permutation tests, $F(x|\boldsymbol{\theta}_1)$ and $F(x|\boldsymbol{\theta}_0)$ can be the distribution functions of either discrete or continuous random variables. If we denote the conditional mean and variance of T_N^P, given H_0, $\mathbf{X}=(X_1,\ldots,X_N)^{\mathrm{T}}$ and m, by $M_0(\mathbf{X},m)$ and $V_0(\mathbf{X},m)$, then Weinberg and Lagakos [7] show that under H_1 and regularity conditions

$$\frac{T_N^P-M_0(\mathbf{X},m)}{\sqrt{\{V_0(\mathbf{X},m)\}}}\xrightarrow{\mathscr{L}}N(\xi,1)\quad\text{as}\quad N\to\infty$$

where

$$\xi = -\frac{\sqrt{\{\lambda(1-\lambda)\}}\sum_{p=1}^{P}\delta_p \int_{-\infty}^{\infty} a(x)\frac{\partial \ln f(x|\boldsymbol{\theta}_0)}{\partial \theta_{0p}}\, \mathrm{d}F(x|\boldsymbol{\theta}_0)}{\sqrt{[\mathrm{var}\{a(X)|\boldsymbol{\theta}_0\}]}} \tag{2}$$

$f(.)$ is the PDF or PMF of $F(.)$ and $\mathrm{var}\{a(X)|\boldsymbol{\theta}_0\}$ is the variance of $a(X)$ under H_0.

For a given setting, the asymptotic relative efficiency of two rank or permutation tests is then given by the squared ratio of their non-centrality parameters ξ.

4. MODEL FOR REPEATED MEASURES DATA

Suppose the outcomes in group g for individual i follow a linear growth curve model, with individuals following one of four types of profiles, say PF_1, PF_2, PF_3 and PF_4, with probabilities P_{1g}, P_{2g}, P_{3g}, and P_{4g}. If Y_{gik} denotes the outcome in group g for individual i at time t_k, then we assume

$$Y_{gik} = \begin{cases} (\beta_0 + b_{0i}) + (\beta_{1g} + b_i)t_k + e_{gik} & \text{w.p. } P_{1g} \\ (\beta_0 + b_{0i}) + (\beta_{2g} + b_i)\{t_k + (\eta - 1)(t_k - t_{\mathrm{N}})(I1_{gk} + I2_{gk})\} + e_{gik} & \text{w.p. } P_{2g} \\ (\beta_0 + b_{0i}) + (\beta_{3g} + b_i)[t_k - (t_k - t_{\mathrm{N}})I1_{gk} + & \\ \qquad \{\tau(t_k - t_{\mathrm{C}g}) - (t_k - t_{\mathrm{N}})\}I2_{gk}] + e_{gik} & \text{w.p. } P_{3g} \\ (\beta_0 + b_{0i}) + (\beta_{4g} + b_i)\{t_k - (t_k - t_{\mathrm{N}})(I1_{gk} + I2_{gk})\} + e_{gik} & \text{w.p. } P_{4g} \end{cases}$$

Here t_{N} is a nadir time point, $t_{\mathrm{C}g}$ is a change point occurring after the nadir, η and τ are constants that determine the fixed slope during specific study periods, $I1_{gk} = 1$ if $t_{\mathrm{N}} \leqslant t_k \leqslant t_{\mathrm{C}g}$, 0 otherwise, $I2_{gk} = 1$ if $t_k > t_{\mathrm{C}g}$, 0 otherwise. For profile 1 the expected value of the slope is β_{1g}. For profile 2 the expected value of the slope is β_{2g} before t_{N} and $\beta_{2g}\eta$ after t_{N}. For profile 3 the expected value of the slope is β_{3g} before t_{N}, 0 between t_{N} and $t_{\mathrm{C}g}$, and $\beta_{3g}\tau$ after $t_{\mathrm{C}g}$. For profile 4, the expected value of the slope is β_{4g} before t_{N} and 0 after t_{N}. We also assume that the measurement errors are independently $\mathrm{N}(0, \sigma_g^2)$ and independent of the individual effects $\mathbf{b}_i$ where

$$\mathbf{b}_i = (b_{0i}, b_i) \sim \mathrm{N}(\mathbf{0}, \mathbf{D}) \text{ with } \mathbf{D} = \begin{pmatrix} d_0 & d_{01} \\ d_{10} & d_1 \end{pmatrix}$$

An example of the expected response for each type of profile of a repeated measures response is shown in Figure 4 for one treatment group, based on the above model with the nadir and change time points at weeks 8 and 16, respectively. We use plasma HIV-1 viral load, currently the most common way of assessing the antiviral activity of HIV drugs, for illustration. In profile 1 viral load fluctuates about baseline levels since $\beta_1 = 0$. For profile 2 the mean viral load level decreases to some nadir above the first threshold, L, at time t_{N} and then returns to baseline levels by the end of the study. Here, $\beta_2 < 0$ and $\eta = -0.5$. In profile 3, mean viral load levels decrease below L by the nadir time point, t_{N}, maintains the treatment effect until the change point, t_{C}, and then increases above L by the end of study. Here, $\beta_3 < 0$ and τ is some negative constant. Finally, in profile 4, mean viral load levels decrease below an even lower threshold, L', by the nadir time point, with $\beta_4 < 0$, and maintain this treatment

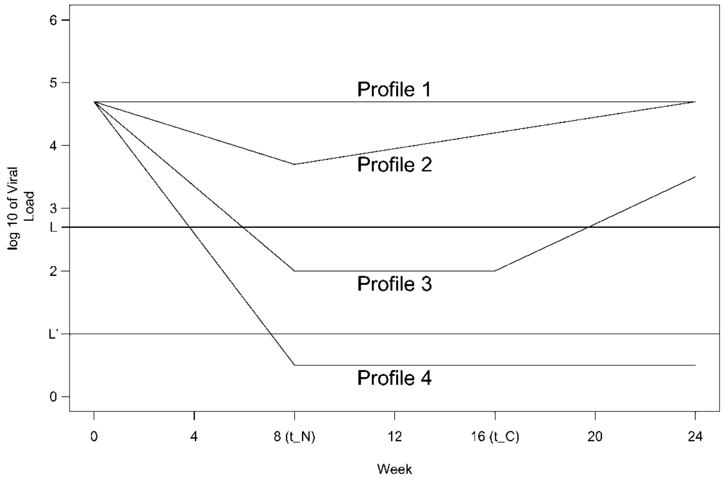

Figure 4. Viral load profiles.

effect throughout the rest of the study. Although this model is very general, in this example we assume that $\beta_1 > \beta_2 > \beta_3 > \beta_4$ so that a sharper initial decline leads to a longer duration of response. Note that the parameter values used for the fixed intercept and slopes in Figure 4 are the same values used in the summary statistic comparisons of Section 6.

For viral load data, the first threshold level L is typically the lower limit of quantification of the laboratory assay. Profile 1 may be a typical response for an individual who does not respond too well to treatment. Profile 2 represents a limited response to treatment in which the individual does not achieve undetectable levels of viral load by the nadir time point, thus the treatment effect is not maintained. Profile 3 also represents a limited response to treatment. The individual does achieve and maintains undetectable levels of viral load until the change time point, at which time the treatment is no longer effective in suppressing viral load, perhaps due to resistance. In profile 4, viral load falls below L by time t_{N} and is maintained at undetectable levels through the end of the study.

This model allows the two treatment groups to differ in fixed population slopes, β_{1g}, β_{2g}, β_{3g} and β_{4g}, the change time point, $t_{\mathrm{C}g}$, the measurement error variance, σ_g^2, and in the probability of observing each of the profile types, P_{1g}, P_{2g}, P_{3g} and P_{4g}. The parameters which differ between treatment groups define the contiguous alternative.

5. SUMMARY STATISTICS

There are a wide variety of summary statistics, X, computed from repeated measures data, which differ in type (discrete or continuous), the amount of information incorporated, clinical relevance and general complexity. We will examine several discrete and continuous type summary statistics which are used in many settings, including AIDS/HIV clinical trials. We assume that measurements are taken at times $t_0 \leqslant t_1 \leqslant \cdots \leqslant t_K$, where t_0 and t_K denote the baseline value and final value, respectively. We consider the following summary statistics:

1. $X=1$ if $Y_K<L$, 0 otherwise, where L is some threshold value (BELOW).
2. $X=1$ if $Y_K<L$ and $Y_{K-1}<L$, 0 otherwise (BELOW2).
3. X is the 'duration of response', or the number of consecutive Y_k at the end of the study which fall below some threshold, L, that is, $X=0$ if $Y_K\geqslant L$, $X=1$ if $Y_K<L$ and $Y_{K-1}\geqslant L$ etc. (DURATION).
4. X is a score variable (SCORE) calculated at the last time point so that

   $$X=\begin{cases}0 & \text{if } Y_K-Y_0>c \text{ and } Y_K>L\\ 1 & \text{if } Y_K-Y_0\leqslant c \text{ and } Y_K>L\\ 2 & \text{if } Y_K\leqslant L\end{cases}$$

   where c is some constant.
5. $X=Y_K^*-Y_0$ (CHANGE).
6. $X=0.5\sum_{k=0}^{K-1}(t_{k+1}-t_k)(Y_{k+1}^*+Y_k^*)=$area under the curve (AUC).
7. $X=0.5\sum_{k=0}^{K-1}(t_{k+1}-t_k)(Y_{k+1}^*-Y_0+Y_k^*-Y_0)=$area under the curve minus baseline(AUCMB).

For summary statistics 5 to 7, $Y_k^*=Y_k$ if $Y_k>L$ and $Y_k^*=C$ if $Y_k\leqslant L$, where C is some constant. That is, if a post-baseline outcome falls below the threshold L, it is replaced by some constant C.

The first four summary statistics are discrete in nature. BELOW considers whether or not the individual has dropped below a threshold at the end of the study. In HIV/AIDS or hepatitis C clinical trials, where viral load is an outcome of considerable interest, this threshold is often a lower limit of quantification, below which an individual's viral load is considered to be undetectable. Alternatively, this threshold may represent an accepted level below which an individual is considered to be in 'good health', for example, total cholesterol level less than 200 mg/dl in clinical trials involving cholesterol lowering agents. The second summary statistic (BELOW2) considers an individual to have truly dropped below the threshold at end of study only if the confirmatory value at time t_{K-1} is also below the threshold. DURATION is aimed at determining the 'duration of response', or how long the treatment effect is maintained at the end of the study. SCORE emphasizes the importance of falling below the threshold value, but considers substantial decreases from baseline (greater than c units), which do not achieve the threshold, to be meaningful as well. Summary statistics 5 to 7 (CHANGE, AUC and AUCMB) are all commonly used continuous metrics. Note that AUCMB adjusts the usual area under the curve for each individual's baseline value, and thus is popular in studies where the baseline variability in marker levels are substantial.

6. ARE OF SUMMARY STATISTICS

In this section we examine the asymptotic relative efficiency of summary statistics, using an AIDS clinical trial for illustration. In Section 6.1 we present five scenarios describing different types of treatment difference, based on the repeated measures model presented in Section 4. We discuss the choice of fixed parameter values in Section 6.2, and present asymptotic relative efficiency (ARE) comparing the summary statistics described in Section 5 for the various scenarios in Section 6.3. For simplicity of presentation, we present results for studies where measurements are taken at baseline and at weeks 4, 8, 12, 16, 20 and 24. The time units themselves are arbitrary but we use these as they reflect common setting for AIDS clinical trials. We also assume that if a post-baseline outcome falls below a lower limit of quantification threshold, L, it is replaced by some constant, C, usually taken to be equal to L, $L/2$ or 0.

6.1. Treatment difference scenarios

The following treatment difference scenarios commonly occur in many disease settings, including the comparison of repeated measures of viral load data in HIV/AIDS clinical trials. In all settings we consider decreases in the outcome to be beneficial. To assess the effect of measurement error variability, each of the following scenarios were examined under equal measurement error variability between the two groups ($\sigma_0^2 = \sigma_1^2$), a small shift in variability ($\sigma_1^2 = 1.5\sigma_0^2$) and a large shift in variability ($\sigma_1^2 = 2\sigma_0^2$).

In scenario 1 the 'better' treatment (group 1) causes a shift to the next profile. For example, a 5 per cent shift would cause 5 per cent of individuals to shift from profile 1 to profile 2, profile 2 to profile 3 and profile 3 to profile 4. For scenario 1.1, we assume that $(P_{10}, P_{20}, P_{30}, P_{40}) = (0.35, 0.35, 0.15, 0.15)$. Here a small, medium or large shift (5, 10 or 15 per cent) corresponds to $(P_{11}, P_{21}, P_{31}, P_{41})$ equal to (0.30, 0.35, 0.15, 0.20), (0.25, 0.35, 0.15, 0.25), or (0.20, 0.35, 0.15, 0.30), respectively. For scenario 1.2 we assume that $(P_{10}, P_{20}, P_{30}, P_{40}) =$ (0.25, 0.25, 0.25, 0.25). Here a small, medium or large shift corresponds to $(P_{11}, P_{21}, P_{31}, P_{41})$ equal to (0.20, 0.25, 0.25, 0.30), (0.15, 0.25, 0.25, 0.35) or (0.10, 0.25, 0.25, 0.40), respectively. For this scenario, $T_{C0} = T_{C1} = 16$.

In scenario 2, group 1 causes a shift from profiles 1 and 2 to profiles 3 and 4. For example, a 5 per cent shift would cause 5 per cent of individuals to shift from profile 1 to profile 3 and from profile 2 to profile 4. For scenario 2.1, we assume that $(P_{10}, P_{20}, P_{30}, P_{40}) = (0.35, 0.35, 0.15, 0.15)$. Here a small, medium or large shift corresponds to $(P_{11}, P_{21}, P_{31}, P_{41})$ equal to $(0.30, 0.30, 0.20, 0.20)$, $(0.25, 0.25, 0.25, 0.25)$ or $(0.20, 0.20, 0.30, 0.30)$, respectively. For scenario 2.2 we assume that $(P_{10}, P_{20}, P_{30}, P_{40}) = (0.25, 0.25, 0.25, 0.25)$. Here a small, medium or large shift corresponds to $(P_{11}, P_{21}, P_{31}, P_{41})$ equal to $(0.20, 0.20, 0.30, 0.30)$, $(0.15, 0.15, 0.35, 0.35)$ or $(0.10, 0.10, 0.40, 0.40)$, respectively. As in scenario 1, $T_{C0} = T_{C1} = 16$.

In scenario 3 the proportion of individuals following each of the four profiles is the same in the two treatment groups, but the treatment groups differ in 'duration of response', that is, where the change time points (T_{C0} and T_{C1}) occur. Each of the following scenarios are examined under a small, medium and large duration shift where a small duration shift corresponds to $T_{C0} = 12$ and $T_{C1} = 16$, a medium duration shift corresponds to $T_{C0} = 12$ and $T_{C1} = 20$, and a large duration shift corresponds to $T_{C0} = 12$ and $T_{C1} = 24$. Here we assume that $(P_{10}, P_{20}, P_{30}, P_{40}) = (P_{11}, P_{21}, P_{31}, P_{41}) = (P_1, P_2, P_3, P_4)$. For scenarios 3.1, 3.2 and 3.3 we assume that (P_1, P_2, P_3, P_4) is equal to $(0.25, 0.25, 0.25, 0.25)$, $(0.20, 0.20, 0.40, 0.20)$, and $(0.15, 0.15, 0.55, 0.15)$, respectively.

In scenario 4, group 1 has a longer duration of response and also has more individuals falling below a threshold, L, by the nadir time point. Each of the scenarios is examined under small, medium and large duration shifts as defined in scenario 3. Here we assume that $(P_{10}, P_{20}, P_{30}, P_{40}) = (0.25, 0.25, 0.25, 0.25)$. Scenarios 4.1, 4.2 and 4.3 correspond to a small, medium or large shift in the proportion of individuals falling below L at the nadir time point, that is, $(P_{11}, P_{21}, P_{31}, P_{41})$ equal to $(0.20, 0.20, 0.30, 0.30)$, $(0.15, 0.15, 0.35, 0.35)$, and $(0.10, 0.10, 0.40, 0.40)$, respectively.

In scenario 5, group 1 has more individuals falling below the threshold L by the nadir time point, but a shorter duration of response, so that it is not clear which treatment is preferable. As in scenarios 3 and 4, each of the following scenarios are examined under a small, medium and large duration shift, however here group 0 has the longer duration so that a small duration shift corresponds to $T_{C0} = 16$ and $T_{C1} = 12$, a medium duration shift corresponds to $T_{C0} = 20$ and $T_{C1} = 12$, and a large duration shift corresponds to $T_{C0} = 24$ and $T_{C1} = 12$. Scenarios 5.1, 5.2 and 5.3 correspond to a small, medium or large shift in the proportion of individuals falling below L (with more individuals in group 1 falling below L at the nadir), as defined for scenarios 4.1, 4.2 and 4.3, respectively.

6.2. *Choice of fixed parameter values*

The following fixed parameter values were used for the scenarios examined in Section 6.3 and were taken to be typical of many HIV trials examining viral load (measured on the $\log_{10}$ scale). We assume that $\lambda = 0.5$, that is, equal allocation of individuals between treatment groups, $c = -1$ is the decline from baseline used in SCORE, $L = 2.7$ is the lower limit of quantification of viral load (500 copies/mL), and $L' = 1$ is a second lower threshold for viral load (10 copies/mL). It is assume that if viral load decreases below L' then, apart from measurement error, the treatment effect will be maintained through the end of study. We assume that $t_N = 8$ so that the nadir time point occurs one-third of the way through the study for both treatment groups. For the intercept and profile specific slopes we assume that $\beta_0 = 4.7$ (fixed intercept corresponding to 50000 copies/mL), $\beta_{10} = \beta_{11} = 0$ (fixed slope for profile 1), so that on average viral load will fluctuate about baseline levels, $\beta_{20} = \beta_{21} = -0.125$ (fixed slope for profile 2) which corresponds to a 1 log decrease by the nadir time point on average, $\beta_{30} = \beta_{31} = -0.3375$ (fixed slope for profile 3) which brings viral load down to 2 logs (100 copies/mL) on average by the nadir, that is, on average viral load is between L' and L by the nadir, and $\beta_{40} = \beta_{41} = -0.525$ (fixed slope for profile 4) so that viral load decreases to 0.5 logs by the nadir on average, that is, on average, viral load is below the lower threshold L' by the nadir. The constants which determine the fixed slope during specific study periods are $\eta = -t_N/(t_K - t_N)$, which ensures that for profile 2, on average, viral load will return to baseline levels by the end of the study, and $\tau = \frac{L + 2\sigma_0 - (\beta_0 + \beta_{30} t_N)}{(t_K - t_{C0})\beta_{30}}$ if $t_{C0} \leqslant t_{C1}$, or $\tau = \frac{L + 2\sigma_1 - (\beta_0 + \beta_{31} t_N)}{(t_K - t_{C1})\beta_{31}}$ if $t_{C0} > t_{C1}$, which ensures that for profile 3, on average, viral load will return to two standard deviations above the lower limit of quantification by end of study in the group with the shorter duration of response. Variability estimates are $d_0 = 0.35$, so that only 2.5 per cent of true viral load values of individuals following profile 3 are above L between t_N and t_C, $d_{01} = d_1 = 0$ (that is, assuming a random intercept only model), and $\sigma_0 = 0.26$, an estimate of measurement error variability in group 0 from Paxton *et al.* [11]. The choice of the other parameters, that is, t_{C0}, t_{C1}, σ_1, P_{10}, P_{11}, P_{20}, P_{21}, P_{30} and P_{41}, depends on the scenario of interest.

Figure 5 contains representative plots of mean viral load by treatment group for each scenario. In general, the treatment difference for scenarios 1 and 2 is constant after the initial decline. Scenario 3 shows a delayed treatment effect. In scenario 4, the treatment effect increases over time, while in scenario 5, the treatment effect is transient.

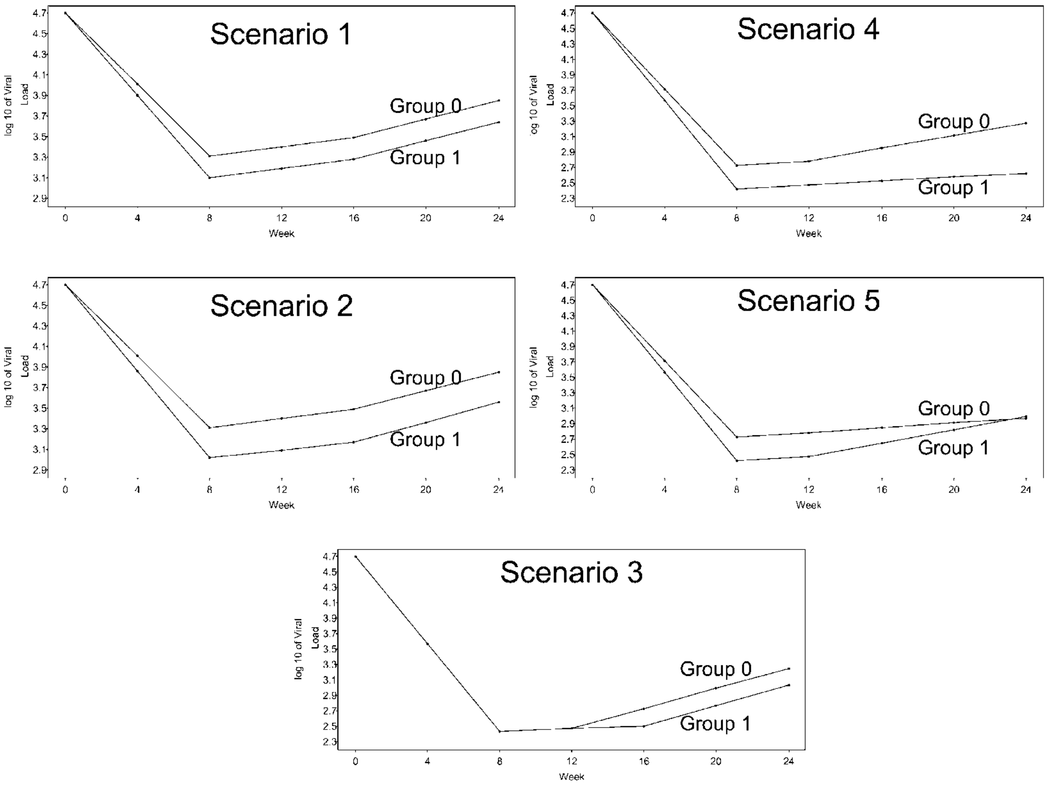

Figure 5. Mean viral load by treatment group.

6.3. *Summary statistic comparisons*

Asymptotic relative efficiencies comparing the summary statistics can be found in Tables I to V for the treatment difference scenarios presented in Section 6.1 assuming that values below the lower limit of quantification, L, are equal to L (that is, $C=L$). In Table VI we provide the probability of falling below L by the nadir, from the nadir to the end of study and at the end of study for each treatment group and scenario. In these tables, permutation tests are used for all discrete summary statistics, and based on untransformed outcomes (that is, $a(X_i)=X_i$), while the Wilcoxon test (that is, $a_N(R_{Ni})=R_{Ni}$), is used for all continuous summary statistics. More details regarding the calculation of the non-centrality parameters can be found in the Appendix. Shifts in variability had little effect on these probabilities therefore they are only presented for the case of no variability shift. More details about these settings and others are available from the authors.

In scenarios 1.1 and 1.2 (see Table I) the treatment effect is a shift in the proportion of individuals following each profile from one profile to the next profile. The relative performance depended very little on the size of the proportion shift, and so only the results for a small proportion shift are presented. For this scenario, group 1 has more undetectable values at the nadir, from the nadir to the end and especially at the end of the study (see Table VI). For

Table I. ARE(SS 1: SS 2) for scenario 1: shift to next profile (small proportion shift).

Scenario	Variability shift		SS 2					
		SS 1	BELOW	BELOW2	DUR.	SCORE	CHANGE	AUC
1.1	none	BELOW2	1.01	1				
		DURATION	1.00	0.99	1			
		SCORE	0.97	0.96	0.97	1		
		CHANGE	0.50	0.49	0.50	0.51	1	
		AUC	0.60	0.60	0.60	0.62	1.20	1
		AUCMB	0.72	0.72	0.73	0.75	1.45	1.21
	small	BELOW2	0.97	1				
		DURATION	0.95	0.99	1			
		SCORE	0.92	0.95	0.96	1		
		CHANGE	0.49	0.50	0.51	0.53	1	
		AUC	0.58	0.60	0.61	0.64	1.21	1
		AUCMB	0.72	0.75	0.76	0.79	1.49	1.23
	large	BELOW2	0.94	1				
		DURATION	0.92	0.98	1			
		SCORE	0.88	0.93	0.95	1		
		CHANGE	0.48	0.51	0.52	0.54	1	
		AUC	0.57	0.61	0.62	0.66	1.21	1
		AUCMB	0.72	0.77	0.78	0.82	1.51	1.25
1.2	none	BELOW2	1.00	1				
		DURATION	0.99	0.99	1			
		SCORE	1.09	1.08	1.10	1		
		CHANGE	0.76	0.76	0.77	0.70	1	
		AUC	0.81	0.80	0.81	0.74	1.06	1
		AUCMB	0.90	0.90	0.91	0.83	1.19	1.12
	small	BELOW2	0.95	1				
		DURATION	0.93	0.98	1			
		SCORE	0.91	0.96	0.98	1		
		CHANGE	0.74	0.78	0.79	0.81	1	
		AUC	0.78	0.82	0.83	0.85	1.05	1
		AUCMB	0.88	0.92	0.94	0.96	1.19	1.13
	large	BELOW2	0.91	1				
		DURATION	0.89	0.98	1			
		SCORE	0.79	0.87	0.89	1		
		CHANGE	0.73	0.80	0.82	0.92	1	
		AUC	0.76	0.83	0.85	0.96	1.04	1
		AUCMB	0.86	0.95	0.97	1.09	1.19	1.14

both of these scenarios all discrete summary statistics perform well relative to the continuous summary statistics. The summary statistic BELOW is best in scenario 1.1 where in group 0 we expect 30 per cent of individuals to fall below L at the nadir and 18 per cent of individuals to fall below L by the end of study (see Table VI). For this scenario the shift in variability causes a small loss of efficiency of the discrete summary statistics BELOW2, DURATION, and SCORE relative to BELOW. For scenario 1.2, where in group 0 48 per cent of individuals are expected to fall below L by the nadir and 30 per cent of individuals to fall below L at end of study, the variability shifts cause a more substantial decline in the performance of BELOW2, DURATION and SCORE, particularly the SCORE summary

Table II. ARE(SS 1: SS 2) for scenario 2: shift from upper to lower profiles (small proportion shift).

Scenario	Variability shift		SS 2					
		SS 1	BELOW	BELOW2	DUR.	SCORE	CHANGE	AUC
2.1	none	BELOW2	1.00	1				
		DURATION	0.99	0.99	1			
		SCORE	1.62	1.63	1.64	1		
		CHANGE	1.21	1.21	1.22	0.74	1	
		AUC	1.13	1.13	1.14	0.70	0.94	1
		AUCMB	1.16	1.16	1.17	0.71	0.96	1.02
	small	BELOW2	0.97	1				
		DURATION	0.96	0.99	1			
		SCORE	1.55	1.60	1.62	1		
		CHANGE	1.17	1.21	1.22	0.76	1	
		AUC	1.10	1.14	1.15	0.71	0.94	1
		AUCMB	1.14	1.18	1.19	0.74	0.98	1.04
	large	BELOW2	0.94	1				
		DURATION	0.93	0.99	1			
		SCORE	1.49	1.58	1.60	1		
		CHANGE	1.14	1.21	1.23	0.77	1	
		AUC	1.08	1.14	1.16	0.72	0.94	1
		AUCMB	1.13	1.20	1.22	0.76	0.99	1.05
2.2	none	BELOW2	1.00	1				
		DURATION	0.99	0.99	1			
		SCORE	1.82	1.83	1.85	1		
		CHANGE	1.71	1.72	1.73	0.94	1	
		AUC	1.60	1.61	1.62	0.88	0.94	1
		AUCMB	1.64	1.64	1.66	0.90	0.96	1.02
	small	BELOW2	0.95	1				
		DURATION	0.94	0.98	1			
		SCORE	1.61	1.69	1.71	1		
		CHANGE	1.65	1.73	1.76	1.03	1	
		AUC	1.54	1.61	1.64	0.96	0.93	1
		AUCMB	1.58	1.66	1.69	0.98	0.96	1.03
	large	BELOW2	0.92	1				
		DURATION	0.90	0.98	1			
		SCORE	1.44	1.57	1.60	1		
		CHANGE	1.59	1.74	1.77	1.11	1	
		AUC	1.49	1.62	1.65	1.03	0.93	1
		AUCMB	1.54	1.67	1.71	1.07	0.96	1.03

statistic. Here, SCORE is best when the treatment groups are similar in variability. However, in general, BELOW is a good choice for scenario 1. Note that AUCMB is the best of the continuous summary statistics with up to 90 per cent efficiency relative to BELOW in scenario 1.2.

In scenarios 2.1 and 2.2 (see Table II) the ‘better’ treatment causes a shift from the two upper profiles to the two lower profiles. For both of these scenarios relative performance is again generally not affected by the size of the proportion shift, therefore results are presented for a small proportion shift. As shown in Table VI, group 1 again has more undetectable values from the nadir to the end of the study, with this increase in undetectable values being more

Table III. ARE(SS 1: SS 2) for scenario 3: longer duration of response (small duration shift).

Scenario	Variability shift	SS 1	SS 2					
			BELOW	BELOW2	DUR.	SCORE	CHANGE	AUC
3.1	none	BELOW2	0.96	1				
		DURATION	0.86	0.90	1			
		SCORE	1.00	1.04	1.16	1		
		CHANGE	1.02	1.06	1.17	1.02	1	
		AUC	0.21	0.22	0.24	0.21	0.20	1
		AUCMB	0.07	0.07	0.08	0.07	0.07	0.33
	small	BELOW2	0.85	1				
		DURATION	0.77	0.90	1			
		SCORE	0.82	0.96	1.07	1		
		CHANGE	0.98	1.14	1.27	1.19	1	
		AUC	0.20	0.24	0.26	0.25	0.21	1
		AUCMB	0.08	0.09	0.10	0.09	0.08	0.38
	large	BELOW2	0.78	1				
		DURATION	0.70	0.90	1			
		SCORE	0.70	0.89	0.99	1		
		CHANGE	0.95	1.22	1.35	1.36	1	
		AUC	0.20	0.26	0.29	0.29	0.21	1
		AUCMB	0.08	0.11	0.12	0.12	0.09	0.41
3.2	none	BELOW2	0.96	1				
		DURATION	0.87	0.90	1			
		SCORE	1.08	1.12	1.24	1		
		CHANGE	1.58	1.64	1.81	1.46	1	
		AUC	0.25	0.26	0.29	0.24	0.16	1
		AUCMB	0.10	0.11	0.12	0.09	0.06	0.40
	small	BELOW2	0.86	1				
		DURATION	0.78	0.90	1			
		SCORE	0.85	0.99	1.09	1		
		CHANGE	1.50	1.73	1.92	1.76	1	
		AUC	0.25	0.29	0.32	0.29	0.16	1
		AUCMB	0.10	0.12	0.13	0.12	0.07	0.42
	large	BELOW2	0.79	1				
		DURATION	0.71	0.90	1			
		SCORE	0.70	0.88	0.98	1		
		CHANGE	1.44	1.82	2.01	2.06	1	
		AUC	0.24	0.30	0.34	0.34	0.17	1
		AUCMB	0.11	0.13	0.15	0.15	0.07	0.44
3.3	none	BELOW2	0.97	1				
		DURATION	0.87	0.90	1			
		SCORE	1.19	1.23	1.36	1		
		CHANGE	2.22	2.30	2.54	1.87	1	
		AUC	0.30	0.31	0.35	0.26	0.14	1
		AUCMB	0.14	0.14	0.16	0.12	0.06	0.46
	small	BELOW2	0.87	1				
		DURATION	0.79	0.90	1			
		SCORE	0.93	1.06	1.18	1		
		CHANGE	2.09	2.41	2.66	2.26	1	
		AUC	0.29	0.34	0.37	0.32	0.14	1
		AUCMB	0.14	0.16	0.17	0.15	0.07	0.47
	large	BELOW2	0.80	1				
		DURATION	0.72	0.91	1			
		SCORE	0.74	0.93	1.03	1		
		CHANGE	2.00	2.51	2.77	1.87	1	
		AUC	0.28	0.35	0.39	0.38	0.14	1
		AUCMB	0.14	0.17	0.19	0.18	0.07	0.49

Table IV. ARE(SS 1: SS 2) for scenario 4: more below *L* and longer duration of response (no variability shift).

Scenario	Duration shift	SS 1	SS 2					
			BELOW	BELOW2	DUR.	SCORE	CHANGE	AUC
4.1	small	BELOW2	0.97	1				
		DURATION	0.93	0.95	1			
		SCORE	1.44	1.48	1.55	1		
		CHANGE	1.39	1.43	1.50	0.97	1	
		AUC	0.84	0.86	0.90	0.58	0.60	1
		AUCMB	0.71	0.73	0.77	0.50	0.51	0.85
	medium	BELOW2	0.97	1				
		DURATION	0.91	0.94	1			
		SCORE	1.30	1.34	1.43	1		
		CHANGE	1.27	1.31	1.40	0.98	1	
		AUC	0.61	0.62	0.67	0.47	0.48	1
		AUCMB	0.45	0.46	0.49	0.35	0.35	0.74
	large	BELOW2	0.97	1				
		DURATION	0.90	0.93	1			
		SCORE	1.23	1.27	1.36	1		
		CHANGE	1.21	1.25	1.35	0.99	1	
		AUC	0.49	0.51	0.55	0.40	0.41	1
		AUCMB	0.33	0.34	0.37	0.27	0.27	0.67
4.2	small	BELOW2	0.98	1				
		DURATION	0.95	0.97	1			
		SCORE	1.57	1.61	1.66	1		
		CHANGE	1.50	1.54	1.59	0.95	1	
		AUC	1.09	1.11	1.15	0.69	0.72	1
		AUCMB	1.01	1.03	1.06	0.64	0.67	0.93
	medium	BELOW2	0.97	1				
		DURATION	0.93	0.95	1			
		SCORE	1.44	1.48	1.55	1		
		CHANGE	1.39	1.43	1.50	0.97	1	
		AUC	0.84	0.86	0.90	0.58	0.60	1
		AUCMB	0.71	0.73	0.77	0.50	0.51	0.85
	large	BELOW2	0.97	1				
		DURATION	0.92	0.95	1			
		SCORE	1.36	1.40	1.48	1		
		CHANGE	1.32	1.36	1.44	0.97	1	
		AUC	0.70	0.72	0.76	0.51	0.53	1
		AUCMB	0.55	0.57	0.60	0.41	0.42	0.79
4.3	small	BELOW2	0.98	1				
		DURATION	0.96	0.98	1			
		SCORE	1.64	1.67	1.71	1		
		CHANGE	1.56	1.59	1.63	0.95	1	
		AUC	1.21	1.24	1.27	0.74	0.78	1
		AUCMB	1.16	1.18	1.21	0.71	0.75	0.96
	medium	BELOW2	0.98	1				
		DURATION	0.94	0.96	1			
		SCORE	1.52	1.56	1.62	1		
		CHANGE	1.46	1.49	1.55	0.96	1	
		AUC	0.99	1.01	1.05	0.65	0.68	1
		AUCMB	0.89	0.91	0.94	0.58	0.61	0.90
	large	BELOW2	0.97	1				
		DURATION	0.93	0.95	1			
		SCORE	1.44	1.48	1.55	1		
		CHANGE	1.39	1.43	1.50	0.97	1	
		AUC	0.84	0.86	0.90	0.58	0.60	1
		AUCMB	0.71	0.73	0.77	0.50	0.51	0.85

Table V. ARE(SS1:SS 2) for scenario 5: more below L but shorter duration of response (no variability shift).

Scenario	Duration shift	SS 1	SS 2					
			BELOW	BELOW2	DUR.	SCORE	CHANGE	AUC
5.1	small	BELOW2	0.79	1				
		DURATION	2.17	2.75	1			
		SCORE	32.87	41.73	15.16	1		
		CHANGE	27.97	35.52	12.90	0.85	1	
		AUC	49.19	62.45	22.69	1.50	1.76	1
		AUCMB	61.77	78.43	28.49	1.88	2.21	1.26
	medium	BELOW2	0.41	1				
		DURATION	0.04	0.10	1			
		SCORE	1.49	3.62	37.63	1		
		CHANGE	5.52	13.44	139.73	3.71	1	
		AUC	7.22	17.59	182.92	4.86	1.31	1
		AUCMB	7.46	18.16	188.79	5.02	1.35	1.03
	large	BELOW2	0.95	1				
		DURATION	0.94	0.98	1			
		SCORE	1.11	1.17	1.18	1		
		CHANGE	0.85	0.89	0.91	0.77	1	
		AUC	0.81	0.85	0.87	0.73	0.95	1
		AUCMB	0.81	0.85	0.87	0.73	0.95	1.00
5.2	small	BELOW2	0.96	1				
		DURATION	1.09	1.14	1			
		SCORE	3.18	3.31	2.91	1		
		CHANGE	2.71	2.82	2.47	0.85	1	
		AUC	3.32	3.46	3.03	1.04	1.23	1
		AUCMB	3.72	3.87	3.40	1.17	1.37	1.12
	medium	BELOW2	1.41	1				
		DURATION	2.01	1.42	1			
		SCORE	5.86	4.15	2.91	1		
		CHANGE	8.17	5.78	4.06	1.39	1	
		AUC	9.03	6.39	4.49	1.54	1.11	1
		AUCMB	9.19	6.50	4.57	1.57	1.13	1.02
	large	BELOW2	0.95	1				
		DURATION	0.94	0.98	1			
		SCORE	1.11	1.17	1.18	1		
		CHANGE	0.85	0.89	0.91	0.77	1	
		AUC	0.81	0.85	0.87	0.73	0.95	1
		AUCMB	0.81	0.85	0.87	0.73	0.95	1.00
5.3	small	BELOW2	0.97	1				
		DURATION	1.04	1.07	1			
		SCORE	2.36	2.43	2.27	1		
		CHANGE	2.01	2.06	1.93	0.85	1	
		AUC	2.23	2.30	2.15	0.95	1.11	1
		AUCMB	2.41	2.49	2.32	1.02	1.20	1.08
	medium	BELOW2	1.15	1				
		DURATION	1.34	1.17	1			
		SCORE	2.68	2.34	2.01	1		
		CHANGE	3.02	2.63	2.26	1.13	1	
		AUC	3.17	2.77	2.37	1.18	1.05	1
		AUCMB	3.21	2.80	2.40	1.20	1.06	1.01
	large	BELOW2	0.95	1				
		DURATION	0.94	0.98	1			
		SCORE	1.11	1.17	1.18	1		
		CHANGE	0.85	0.89	0.91	0.77	1	
		AUC	0.81	0.85	0.87	0.73	0.95	1
		AUCMB	0.81	0.85	0.87	0.73	0.95	1.00

Table VI. Probability of falling below L at the nadir (Nadir), from the nadir to end of study (Nadir-End) and at end of study (End).

	Proportion shift		Group 0			Group 1	
Scenario		Nadir	Nadir-End	End	Nadir	Nadir-End	End
1.1	small	0.30	0.18	0.18	0.35	0.23	0.23
	medium	0.30	0.18	0.18	0.40	0.28	0.28
	large	0.30	0.18	0.18	0.45	0.33	0.33
1.2	small	0.48	0.30	0.30	0.53	0.35	0.35
	medium	0.48	0.30	0.30	0.58	0.40	0.40
	large	0.48	0.30	0.30	0.63	0.45	0.45
2.1	small	0.30	0.18	0.18	0.39	0.24	0.24
	medium	0.30	0.18	0.18	0.48	0.30	0.30
	large	0.30	0.18	0.18	0.57	0.36	0.36
2.2	small	0.48	0.30	0.30	0.57	0.36	0.36
	medium	0.48	0.30	0.30	0.66	0.42	0.42
	large	0.48	0.30	0.30	0.75	0.48	0.48
3.1	small	0.48	0.30	0.30	0.48	0.35	0.36
	medium	0.48	0.30	0.30	0.48	0.41	0.42
	large	0.48	0.30	0.30	0.48	0.44	0.47
3.2	small	0.56	0.28	0.28	0.56	0.37	0.37
	medium	0.56	0.28	0.28	0.56	0.46	0.47
	large	0.56	0.28	0.28	0.56	0.50	0.54
3.3	small	0.63	0.26	0.27	0.63	0.38	0.39
	medium	0.63	0.26	0.27	0.63	0.50	0.52
	large	0.63	0.26	0.27	0.63	0.56	0.62
4.1	small	0.48	0.30	0.30	0.57	0.42	0.43
	medium	0.48	0.30	0.30	0.57	0.49	0.50
	large	0.48	0.30	0.30	0.57	0.52	0.56
4.2	small	0.48	0.30	0.30	0.66	0.49	0.50
	medium	0.48	0.30	0.30	0.66	0.57	0.59
	large	0.48	0.30	0.30	0.66	0.61	0.65
4.3	small	0.48	0.30	0.30	0.75	0.56	0.57
	medium	0.48	0.30	0.30	0.75	0.66	0.67
	large	0.48	0.30	0.30	0.75	0.70	0.74
5.1	small	0.48	0.35	0.36	0.57	0.36	0.36
	medium	0.48	0.41	0.42	0.57	0.36	0.36
	large	0.48	0.44	0.47	0.57	0.36	0.36
5.2	small	0.48	0.35	0.36	0.66	0.42	0.42
	medium	0.48	0.41	0.42	0.66	0.42	0.42
	large	0.48	0.44	0.47	0.66	0.42	0.42
5.3	small	0.48	0.35	0.36	0.75	0.48	0.48
	medium	0.48	0.41	0.42	0.75	0.48	0.48
	large	0.48	0.44	0.47	0.75	0.48	0.48

extreme than in scenario 1. In scenario 2.1, the variability shift has only a small effect. For this scenario SCORE is best at detecting treatment differences and is substantially more efficient than the other summary statistics. The continuous summary statistics are generally more efficient than the discrete summary statistics with the exception of SCORE. For scenario 2.2, the variability shift has a more pronounced effect than in scenario 2.1 with the SCORE summary statistic, in particular, losing some efficiency as the variability shift increases. Although

SCORE, in general, is most efficient for this scenario, the continuous summary statistics have now gained efficiency relative to SCORE, and in some cases, outperforms SCORE when there are variability shifts. For both of these scenarios, AUC and AUCMB perform comparably. Note that here the proportion of individuals falling in profiles 1 to 4 differ between treatment groups (compared to only profiles 1 and 4 in scenario 1). This is likely to account for increased efficiency of both the continuous summary statistics and the SCORE statistic which are more sensitive to shifts across profiles than the other discrete summary statistics.

In Table III we present the ARE comparisons for scenario 3 when there is a small duration shift. In this set of scenarios the two treatment groups differ only in duration of response, with progressively more individuals following profile 3, where duration of response comes into play, as we move from scenario 3.1 to 3.3. Here, the two groups have nearly identical proportions of undetectable values at the nadir, but group 1 has substantially more undetectable values by the end of the study (see Table VI). For all three scenarios the size of the duration shift does not affect the relative performance of the summary statistics when there is no shift in variability. The variability shift does affect relative performance. For example, BELOW2, DURATION and SCORE all lose efficiency relative to BELOW as the variability shift increases, however, the effect of the variability shift declines as the duration shift increases. When there are equal proportions of individuals following each of the four profiles in both groups (scenario 3.1) both BELOW and CHANGE perform well, with SCORE performing comparably when there is no variability shift between treatment groups. As progressively more individuals follow profile 3 (scenarios 3.2 and 3.3) CHANGE becomes more and more efficient relative to the other summary statistics. For all three scenarios, both AUC and AUCMB perform very poorly relative to the other summary statistics, with AUCMB being much less efficient than AUC. Note that both AUC and AUCMB compare treatment groups at all post-baseline time points, however, here the treatments only differ late in the study due to the shift in duration, which accounts for the extreme inefficiency of these summary statistics. Here, AUCMB may be less efficient than AUC due to the incorporation of the baseline value, which only adds noise rather than aiding in the detection treatment differences in this setting. In general, CHANGE appears to be a good choice for scenario 3, most likely due to its sensitivity to the level of viral load at the end of the study, where other statistics (such as BELOW) are only sensitive to viral load levels above or below L at the end of the study.

In scenarios 4.1, 4.2 and 4.3 (see Table IV), group 1 not only has a longer duration of response, but also has more individuals falling below the lower limit of quantification. Here we present results for the case of no variability shift as variability shifts had only a small effect on the relative performance of summary statistics. For this scenario group 1 has more undetectable values at the nadir, from the nadir to the end of study and at the end of the study, with a larger increase in undetectable values compared to other scenarios. Here, for all three scenarios, both SCORE and CHANGE perform well relative to other summary statistics and perform comparably to each other. Both of these summary statistics lose some efficiency (relative to other summary statistics) as the shift in duration increases but gain efficiency as the shift in proportion increases (with group 1 having progressively more individuals falling below L in scenarios 4.1 to 4.3). SCORE again tended to lose efficiency as the variability shift increased. The relative performance of both AUC and AUCMB improves dramatically with the addition of a shift in proportion, compared to only a duration shift in scenarios 3.1 to 3.3. In fact, for a small duration shift, the performances of AUC and AUCMB surpass the performances of BELOW, BELOW2 and DURATION as the shift in proportion increases.

Here, in general, AUC performs better than AUCMB, however the performance of AUCMB improves relative to AUC as the shift in proportion increases. Both SCORE and CHANGE appear to be good choices for this set of scenarios.

In Table V we present the ARE comparison for scenarios 5.1, 5.2 and 5.3 with no variability shift. As in scenario 4, variability shifts do not substantially alter our conclusions regarding the relative performance of summary statistics. For this set of scenarios, group 1 has more individuals falling below the lower limit of quantification but has a shorter duration of response. It is thus unclear which treatment group is preferable. This is further illustrated in Table VI, in that for a portion of these scenarios an individual in group 0 is more likely to stay below quantifiable limits from the nadir to the end of the study and at the end of the study, while for other settings an individual in group 1 is more likely to be below this lower limit. This is the only set of scenarios where a summary statistic's non-centrality parameter may be either positive or negative, leading to unusual fluctuations in ARE. In all three scenarios AUCMB is the best choice when there are small or medium duration shifts, while SCORE is best when there are large duration shifts. AUCMB generally had greater or equal performance compared to AUC. In general, the performance of the discrete summary statistics tends to improve as the duration shift increases while the performance of the continuous summary statistics, particularly AUC and AUCMB, tended to stay constant. The extremely large AREs seen in scenario 5.1 for short duration shifts illustrates the poor performance of BELOW, BELOW2 and DURATION in detecting treatment differences, due to similar proportions of individuals falling below L in each treatment group. The large duration shift corresponds to one group maintaining the treatment effect through the end of the study, which helps explain why SCORE performs well in this situation. However, for the other settings it appears to be preferable to use a summary statistic such as AUCMB which incorporates intermediate values and retains the ability to detect treatment differences even when one group may be better in terms of having more individuals below L, but worse in terms of duration of response.

The performance of the continuous summary statistics CHANGE, AUC and AUCMB are affected by the choice of the constant C used to replace values falling below the lower limit of quantification. We examined ARE comparing the Wilcoxon test for the summary statistic CHANGE when different values of the constant C ($C=0$ versus $C=L$) are used to replace values falling below L. For scenarios 1 and 5, using $C=0$ is generally more efficient; for scenarios 3 and 4, using $C=L$ is more efficient; and for scenario 2 the two constants appear to be equally efficient. Although the two constants are, in general, not equally efficient, replacing $C=L$ with $C=0$ in Section 5.3 would not have substantially altered the conclusions regarding the relative performance of summary statistics. The fact that neither constant is superior in all the scenarios supports the use of an intermediate value such as $C=L/2$. Further details are available upon request.

7. EXAMPLES REVISITED

We now return to the examples presented in Section 2. First consider the subgroup of patients that were not taking a protease inhibitor at the time of randomization in the study by Kahn *et al.* where the treatment effect occurs early and is maintained for the 24-week study period (see Figure 1). Treatment group comparisons based on the summary statistics described in Section 5 can be found in Table VII. The number (percentage) of individuals falling into

Table VII. Treatment group comparisons for Gilead Science Study 408, patients not taking a protease inhibitor at time of randomization.

Summary	Placebo ($n = 110$)	Adefovir ($n = 89$)	$\|Z\|$	P-value
BELOW	8 (7.3)	13 (14.6)	1.67	0.0949
BELOW2	4 (3.6)	9 (10.1)	1.83	0.0667
DURATION				
0	102 (92.7)	76 (85.4)	2.08	0.0374
1	4 (3.6)	4 (4.5)		
2	3 (2.7)	4 (4.5)		
3	0 (0.0)	0 (0.0)		
4	1 (0.9)	5 (5.6)		
SCORE				
0	99 (90.0)	68 (76.4)	2.30	0.0216
1	3 (2.7)	8 (9.0)		
2	8 (7.3)	13 (14.6)		
CHANGE	−0.04 (0.57)	−0.47 (0.68)	4.81	< 0.0001
AUC	99 (11)	93 (14)	3.62	0.0003
AUCMB	−1.2 (8.3)	−9.7 (10.8)	6.10	< 0.0001

specified categories is presented for the discrete summary statistics, while the mean (SD) is presented for the continuous summary statistics. Note that the continuous summary statistics result in much more significant test statistics than the discrete summary statistics. This treatment effect is similar to scenario 2 (see Figure 5) where the ARE results indicated that the continuous summary statistics and the SCORE statistic should do well. In fact, SCORE also performs reasonably well in this example. However, viral load values in the Adefovir group are more highly variable than in the placebo group, which may account for the relative lower significance of the test based on SCORE, which loses power with shifts in variability.

We next examine the treatment effect on viral load in the subgroup of patients taking a protease inhibitor at the time of randomization. Recall that for this example the treatment effect occurs primarily towards the end of the study, with the largest difference seen at week 24 (see Figure 2). Treatment group comparisons can be found in Table VIII. Here, the discrete summary statistics resulted in more significant test statistics than the continuous summary statistics, with BELOW, SCORE and DURATION all resulting in highly significant test statistics ($p=0.0006$, $p=0.0011$ and $p=0.0032$, respectively). Both AUC and AUCMB have little power to detect treatment differences in this setting because they include comparisons at time points where there is not yet a difference. This treatment effect in this example is most similar to scenario 3 (see Figure 5). It is not surprising that both AUC and AUCMB do not perform well here, yet the magnitude of the difference in p-values for tests corresponding to different summary statistics is surprising. Based on the ARE results, it is not completely clear why BELOW, DURATION and SCORE would be expected to do well in this setting, which appears to be a cross between scenarios 3 and 4.

Finally we consider Study ACTG 241. Here there is a short-term treatment effect by week 4, which diminishes by the end of the study period (see Figure 3). Treatment group comparisons can be found in Table IX. As would be expected, summary statistics such as BELOW and BELOW2 do not do well in this setting because they do not detect the early treatment differences. In this setting AUCMB results in a substantially more significant test statistic

Table VIII. Treatment group comparisons for Gilead Science Study 408, patients taking a protease inhibitor at time of randomization.

Summary	Placebo ($n = 71$)	Adefovir ($n = 72$)	$\|Z\|$	P-value
BELOW	0 (0.0)	11 (15.3)	3.42	0.0006
BELOW2	0 (0.0)	5 (6.9)	2.25	0.0243
DURATION				
0	71 (100.0)	61 (84.7)	2.94	0.0032
1	0 (0.0)	6 (8.3)		
2	0 (0.0)	2 (2.8)		
3	0 (0.0)	2 (2.8)		
4	0 (0.0)	1 (1.4)		
SCORE				
0	68 (95.8)	58 (80.6)	3.26	0.0011
1	3 (4.2)	3 (4.2)		
2	0 (0.0)	11 (15.3)		
CHANGE	−0.10 (0.43)	−0.31 (0.74)	1.16	0.2449
AUC	99 (12)	98 (12)	0.40	0.6864
AUCMB	−3.3 (7.5)	−4.5 (9.5)	0.38	0.7043

Table IX. Treatment group comparisons for ACTG 241.

Summary	Double therapy ($n = 100$)	Triple therapy ($n = 103$)	$\|Z\|$	P-value
BELOW	7 (7.0)	11 (10.7)	0.92	0.3577
BELOW2	7 (7.0)	9 (8.7)	0.46	0.6468
DURATION				
0	93 (93.0)	92 (89.3)	1.07	0.2862
1	0 (0.0)	2 (1.9)		
2	4 (4.0)	1 (1.0)		
3	0 (0.0)	1 (1.0)		
4	3 (3.0)	7 (6.8)		
SCORE				
0	88 (88.0)	85 (82.5)	1.09	0.2768
1	5 (5.0)	7 (6.8)		
2	7 (7.0)	11 (10.7)		
CHANGE	−0.13 (0.57)	−0.28 (0.74)	1.41	0.1582
AUC	106 (23)	101 (25)	1.21	0.2270
AUCMB	−5.4 (10.5)	−11.3 (14.4)	3.01	0.0027

than other summary statistics ($p=0.0027$). The treatment effect in this example is similar to scenario 5 (see Figure 5) where the treatment effect is transient. Based on the ARE results for this scenario we may have expected AUC to perform nearly as well as AUCMB. The apparently greater power for AUCMB can be explained in part by substantial between-subject variability in baseline viral load levels and the treatment effect not depending strongly on these baseline levels. Thus, a summary statistic such as AUCMB more efficiently contols for the between-subject variability than a summary statistic such as AUC.

8. DISCUSSION

The efficiency of a summary statistic for comparing two groups depends in a complex way on several factors, including the underlying longitudinal model, the treatment effect, the number and timing of measurements, measurement error, the threshold of interest, the presence of a lower limit of quantification and the constant C used to replace values falling below a lower limit of quantification. The choice of summary statistic should also depend on clinical considerations of what types of effects are important. With regard to our HIV trial examples, there is currently great uncertainty regarding the clinical importance of different aspects of an individual's viral load profile. For example, is the only important impact of treatment to lower viral load below quantifiable limits or are other results also clinically meaningful? Questions such as this help determine a clinically meaningful summary statistic, however, in general, the goal should be to choose a clinically relevant summary statistic which is also fairly efficient.

Based on the settings we have examined, BELOW2 was usually less efficient than BELOW and DURATION was usually less efficient than BELOW2, although the difference in efficiency was generally not substantial. Thus, BELOW would generally be preferred to BELOW2 if a binary measure of response was desired. This was a bit surprising because both BELOW2 and DURATION utilize more measurements than BELOW, yet they are also influenced by measurement error at each time point, so that looking at these additional time points may often add more noise than information about treatment differences. Not surprisingly, BELOW was most efficient when the main treatment difference was simply characterized by more individuals falling below a lower threshold in one group, as in scenario 1. In other situations other summary statistics can be substantially more efficient than BELOW.

The SCORE summary statistic can be viewed as a sort of hybrid between CHANGE and BELOW, and may therefore be expected to perform well in a variety of situations. Our evaluations show that SCORE, although not always optimal, performs fairly well in many of the situations examined, particularly when the two treatment groups were similar with respect to measurement error. SCORE, in general, tended to lose efficiency with an increasing shift in variability. CHANGE from baseline to the end of the study also performed fairly well in several situations, with the exception of scenario 1, where the treatment difference was clearly focused on whether the outcome was above or below a threshold.

The AUC and AUCMB summary statistics appear to be best in situations where neither treatment group is dominant for the entire study period. As mentioned previously these statistics use intermediate values and retain the ability to detect treatment differences even when one group may be better in some respects (for example, more individuals below L) but worse in other respects (for example, shorter duration of response). As illustrated in scenario 3, AUC and AUCMB may also perform extremely poorly, particularly when the treatment difference is focused on a specific portion of the response profile.

We have not examined the consequences of missing values of the response process in our comparison of summary statistics. One consideration is how summary statistics would be defined in the presence of unanticipated missing data. Clearly, missing data can have a greater impact on some summary statistics than others; for example, missing values prior to the final observation time have no impact on BELOW but will affect AUC. Another consideration is the effect of non-informatively missing observations on the relative efficiency of the summary statistics. Finally, if observations are informatively missing, the validity of the tests considered in this paper may be compromised. These issues are all relevant to the choice of summary statistics, but are beyond the scope of this paper.

In much of this paper we have discussed settings in the context of HIV trials that compare groups with respect to viral load. However, the settings and scenarios considered occur in other disease settings, especially when the repeated measures process is based on a laboratory marker and when treatment effects may be transient. In choosing a summary statistic in other settings, one should consider the results presented above as well as the scenarios that seem, *a priori*, most plausible.

Finally, we note that our comparisons are based on asymptotic results, however simulations have supplied evidence of good small sample accuracy of the results in Section 3. For further details see Weinberg and Lagakos [6, 7].

APPENDIX

All calculations were performed using Maple V release 5. A permutation test based on untransformed outcomes (that is, $a(X_i)=X_i$) was used for all discrete summary statistics. Here, the non-centrality parameter for the permutation test in (2) reduces to

$$\xi = -\frac{\sqrt{\{\lambda(1-\lambda)\}}\sum_{p=1}^{P}\delta_p\sum_x x\frac{\partial \Pr(X=x|\boldsymbol{\theta}_0)}{\partial\theta_{0p}}}{\sqrt{\{\mathrm{var}(X|\boldsymbol{\theta}_0)\}}}$$

where $\Pr(X=x|\boldsymbol{\theta}_0)$ is the probability that the summary statistic is equal to the value x in group 0, and P is the total number of parameters in the contiguous alternative.

Note that

$$\begin{aligned}\Pr(X=x|\boldsymbol{\theta}_0) &= \Pr(X=x|\boldsymbol{\theta}_0,\ \mathrm{PF}_1)P_{10}+\Pr(X=x|\boldsymbol{\theta}_0,\ \mathrm{PF}_2)P_{20}+\Pr(X=x|\boldsymbol{\theta}_0,\ \mathrm{PF}_3)P_{30}\\ &\quad+\Pr(X=x|\boldsymbol{\theta}_0,\ \mathrm{PF}_4)(1-P_{10}-P_{20}-P_{30})\end{aligned}$$

For the BELOW and SCORE summary statistics, the quantities on the right hand side can be obtained directly. For example, for the BELOW summary statistic

$$\Pr(X=1|\boldsymbol{\theta}_0,\ \mathrm{PF}_1)=\Pr(Y_K\leqslant L|\boldsymbol{\theta}_0,\ \mathrm{PF}_1)=\Phi\left(\frac{L-M1_K}{\sqrt{(V1_K)}}\right)$$

while for SCORE

$$\begin{aligned}\Pr(X=1|\boldsymbol{\theta}_0,\ \mathrm{PF}_1) &= \Pr(Y_K-Y_0\leqslant -1 \text{ and } Y_K>L|\boldsymbol{\theta}_0,\ \mathrm{PF}_1)\\ &= \Pr(Y_K-Y_0\leqslant -1\,|\,Y_K>L,\boldsymbol{\theta}_0,\ \mathrm{PF}_1)\Pr(Y_K>L|\boldsymbol{\theta}_0,\ \mathrm{PF}_1)\\ &= \Phi\left(\frac{-1-Mchg1}{\sqrt{(Vchg1)}}\right)\left\{1-\Phi\left(\frac{L-M1_K}{\sqrt{(V1_K)}}\right)\right\}\end{aligned}$$

where Φ is the standard normal CDF, $M1_K$ and $V1_K$ represent the mean and variance of Y_K in group 0 for profile 1(PF_1) and $Mchg1$ and $Vchg1$ are the mean and variance of change from baseline to end of study in group 0 for profile 1. The derivatives, $\partial\Pr(X=1|\boldsymbol{\theta})/\partial\theta_{0p}$, $p=1,\ldots,P$ for these expressions are easily computed.

To find $P(X=x|\mathbf{b},\boldsymbol{\theta}_0)$ for BELOW2 and DURATION, we first condition on the random effect $\mathbf{b}$, obtaining

$$\begin{aligned}\Pr(X=x|\mathbf{b},\boldsymbol{\theta}_0) &= \Pr(X=x|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_1)P_{10}+\Pr(X=x|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_2)P_{20}\\ &\quad+\Pr(X=x|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_3)P_{30}+\Pr(X=x|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_4)(1-P_{10}-P_{20}-P_{30})\end{aligned}$$

Each conditional probability on the right hand side can now be computed. For example, for BELOW2 when $X=1$ we have

$$\begin{aligned}\Pr(X=1|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_1) &= \Pr(Y_K\leqslant L \text{ and } Y_{K-1}\leqslant L|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_1)\\ &= \Pr(Y_K\leqslant L|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_1)\Pr(Y_{K-1}\leqslant L|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_1)\\ &= \Phi\left(\frac{L-\mathrm{CM1}_K}{\sigma_0}\right)\Phi\left(\frac{L-\mathrm{CM1}_{K-1}}{\sigma_0}\right)\end{aligned}$$

where $\mathrm{CM1}_K$ and $\mathrm{CM1}_{K-1}$ are the conditional means of Y_K and Y_{K-1}, respectively, given the random effects $\mathbf{b}$, group 0 and profile 1. Thus

$$\Pr(X=1|\boldsymbol{\theta}_0,\mathrm{PF}_1)=\int_{\mathbf{b}}\int \Pr(X=1|\mathbf{b},\boldsymbol{\theta}_0,\mathrm{PF}_1)h(\mathbf{b})\,\mathrm{d}\mathbf{b},$$

where $h(\mathbf{b})$ is the bivariate normal density of the random effects $\mathbf{b}$. Thus

$$\frac{\partial \Pr(X=1|\boldsymbol{\theta}_0)}{\partial\theta_{0p}}=\int_{\mathbf{b}}\int\frac{\partial}{\partial\theta_{0p}}\left\{\Pr(X=1|\mathbf{b},\boldsymbol{\theta}_0)\right\}h(\mathbf{b})\,\mathrm{d}\mathbf{b}.$$

where we have exchanged the order of the derivative and integral. This integral cannot be solved in closed form. To find the value numerically for a specific value of the parameters, we first find $\partial/\partial\theta_{0p}\{\Pr(X=1|\mathbf{b},\boldsymbol{\theta}_0)\}$, and then perform numerical integration.

The Wilcoxon test was used for all continuous summary statistics. Thus $\varphi'(u)=1$ and $\int_0^1\{\varphi(u)-\bar{\varphi}\}^2\,\mathrm{d}u=1/12$. Letting $x=F^{-1}(u)$, the non-centrality parameter in (1) reduces to

$$\xi=\surd\{12\lambda(1-\lambda)\}\sum_{p=1}^{P}\delta_p\int_{-\infty}^{\infty}\frac{\partial F(x|\boldsymbol{\theta}_0)}{\partial\theta_{0p}}f(x|\boldsymbol{\theta}_0)\,\mathrm{d}x$$

Note that

$$\begin{aligned}F(x|\boldsymbol{\theta}_0)=\Pr(X\leqslant x|\boldsymbol{\theta}_0,\mathrm{PF}_1)P_{10}+\Pr(X\leqslant x|\boldsymbol{\theta}_0,\mathrm{PF}_2)P_{20}+\\ \Pr(X\leqslant x|\boldsymbol{\theta}_0,\mathrm{PF}_3)P_{30}+\Pr(X\leqslant x|\boldsymbol{\theta}_0,\mathrm{PF}_4)(1-P_{10}-P_{20}-P_{30})\end{aligned}$$

Recall that post-baseline outcomes falling below L are replaced by C. Thus, for example, for the CHANGE summary statistic

$$\Pr(X\leqslant x|\boldsymbol{\theta}_0,\mathrm{PF}_1)=\Pr(Y_K^*-Y_0\leqslant x|\boldsymbol{\theta}_0,\mathrm{PF}_1)$$

where

$$Y_K^*=\begin{cases}Y_K & \text{if } Y_K>L\\ C & \text{if } Y_K\leqslant L\end{cases}$$

and hence

$$\begin{aligned}\Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1) &= \Pr(Y_K - Y_0 \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1)\Pr(Y_K > L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\quad + \Pr(C - Y_0 \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1)\Pr(Y_K \leqslant L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &= \Phi\left(\frac{x - Mchg1}{\surd(Vchg1)}\right)\left\{1 - \Phi\left(\frac{L - M1_K}{\surd(V1_K)}\right)\right\}\\ &\quad + \Phi\left\{\frac{x - (C - \beta_0)}{\surd(d_0 + \sigma_0^2)}\right\}\Phi\left(\frac{L - M1_K}{\surd(V1_K)}\right)\end{aligned}$$

where $Mchg1$ and $Vchg1$ are the mean and variance of change from baseline in group 0 for profile 1 when the final value is observed, and $M1_K$ and $V1_K$ are the mean and variance of Y_K in group 0 for profile 1. Note that $C - Y_0 \sim \mathrm{N}(C - \beta_0, d_0 + \sigma_0^2)$ regardless of the profile. The derivatives in the numerator of the non-centrality parameter can be found directly for this summary statistic.

To find the non-centrality parameter for the AUC summary statistic note that

$$X = 0.5\sum_{k=0}^{K-1}(t_{k+1} - t_k)(Y_{k+1}^* + Y_k^*)$$

To simplify calculations, we based AUC only on weeks 0, 8, 16 and 24. To account for the fact that we are using three post-baseline time points rather than six post-baseline time points we assumed that the value at weeks 8, 16 and 24 represented the mean of two measurements, and therefore divided the measurement error variance in half. For notational simplicity, let Y_0, Y_8, Y_{16} and Y_{24} denote the outcomes at weeks 0, 8, 16 and 24, respectively. Then, for example

$$\begin{aligned}&\Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\quad = \Pr\{4(Y_0 + 2Y_8 + 2Y_{16} + Y_{24}) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 > L, Y_{16} > L, Y_{24} > L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 2C + 2Y_{16} + Y_{24}) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 \leqslant L, Y_{16} > L, Y_{24} > L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 2Y_8 + 2C + Y_{24}) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 > L, Y_{16} \leqslant L, Y_{24} > L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 2Y_8 + 2Y_{16} + C) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 > L, Y_{16} > L, Y_{24} \leqslant L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 4C + Y_{24}) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 \leqslant L, Y_{16} \leqslant L, Y_{24} > L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 3C + 2Y_{16}) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 \leqslant L, Y_{16} > L, Y_{24} \leqslant L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 3C + 2Y_8) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 > L, Y_{16} \leqslant L, Y_{24} \leqslant L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\\ &\qquad + \Pr\{4(Y_0 + 5C) \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1\}\Pr(Y_8 \leqslant L, Y_{16} \leqslant L, Y_{24} \leqslant L|\boldsymbol{\theta}_0, \mathrm{PF}_1)\end{aligned}$$

Each of the joint probabilities can be found by first conditioning on the random effects. For example

$$\Pr(Y_8 \leqslant L, Y_{16} \leqslant L, Y_{24} \leqslant L|\boldsymbol{\theta}_0, \mathrm{PF}_1)$$

$$
\begin{aligned}
&= \int_{\mathbf{b}}\int \Pr(Y_8 \leqslant L, Y_{16} \leqslant L, Y_{24} \leqslant L|\mathbf{b}, \boldsymbol{\theta}_0, \mathrm{PF}_1)h(\mathbf{b}) \\
&= \int_{\mathbf{b}}\int \Pr(Y_8 \leqslant L|\mathbf{b}, \boldsymbol{\theta}_0, \mathrm{PF}_1)\Pr(Y_{16} \leqslant L|\mathbf{b}, \boldsymbol{\theta}_0, \mathrm{PF}_1)\Pr(Y_{24} \leqslant L|\mathbf{b}, \boldsymbol{\theta}_0, \mathrm{PF}_1)h(\mathbf{b}) \\
&= \int_{\mathbf{b}}\int \Phi\left(\frac{L-\mathrm{CM1}_8}{\sigma_0}\right)\Phi\left(\frac{L-\mathrm{CM1}_{16}}{\sigma_0}\right)\Phi\left(\frac{L-\mathrm{CM1}_{24}}{\sigma_0}\right)h(\mathbf{b})
\end{aligned}
$$

where $\mathrm{CM1}_8$, $\mathrm{CM1}_{16}$ and $\mathrm{CM1}_{24}$ denote the conditional means (conditioning on the random effects) in group 0 for profile 1, at weeks 8, 16 and 24, respectively. To find the derivatives in the numerator of the non-centrality parameter, note that,

$$
\begin{aligned}
\frac{\partial F(x|\boldsymbol{\theta}_0)}{\partial \theta_{0p}} &= \frac{\partial \Pr(X \leqslant x|\boldsymbol{\theta}_0)}{\partial \theta_{0p}} \\
&= \frac{\partial \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1)}{\partial \theta_{0p}} P_{10} + \frac{\partial \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_2)}{\partial \theta_{0p}} P_{20} \\
&\quad + \frac{\partial \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_3)}{\partial \theta_{0p}} P_{30} + \frac{\partial \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_4)}{\partial \theta_{0p}} (1 - P_{10} - P_{20} - P_{30})
\end{aligned}
$$

where, for example, $\partial \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1)/\partial \theta_{0p}$ is a function of the joint probabilities described above and the derivatives of the joint probabilities which can be found by exchanging differentiation and integration in the same manner described previously. When the contiguous alternative his defined in terms of P_{10}, P_{20} or P_{30} the corresponding derivatives are more easily found, for example

$$
\frac{\partial F(x|\boldsymbol{\theta}_0)}{\partial P_{10}} = \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_1) - \Pr(X \leqslant x|\boldsymbol{\theta}_0, \mathrm{PF}_4)
$$

The non-centrality parameter for area under the curve minus baseline (AUCMB) was found using similar methods to that of AUC. Here

$$
X = 0.5 \sum_{k=0}^{K-1} (t_{k+1} - t_k)(Y^*_{k+1} - Y_0 + Y^*_k - Y_0)
$$

Analogous to the AUC calculations, we based AUCMB only on weeks 0, 8, 16 and 24.

ACKNOWLEDGEMENTS

This research was supported by NIH award AI24643. We thank the referees for their helpful comments.

REFERENCES

1. Pocock SJ. *Clinical Trials*: *A Practical Approach*. Wiley: New York, 1983.
2. Matthews JNS, Altman DG, Campbell MJ, Royston, P. Analysis of serial measurements in medical research. *British Medical Journal* 1990; **300**:230–235.
3. Dawson JD. Comparing treatment groups on the basis of slopes, areas-under-the-curve, and other summary measures. *Drug Information Journal* 1994; **28**:723–732.
4. Dawson JD. Stratification of summary statistic tests according to missing data patterns. *Statistics in Medicine* 1994; **13**:1853–1863.

5. Dawson JD, Lagakos SW. Size and power of two-sample tests of repeated measures data. *Biometrics* 1993; **49**:1022–1032.
6. Weinberg JM, Lagakos SW. Linear rank tests under general alternatives, with application to summary statistics computed from repeated measures data. *Journal of Statistical Planning and Inference* (in press).
7. Weinberg JM, Lagakos SW. Asymptotic behaviour of linear permutation tests under general alternatives, with application to test selection and study design. *Journal of the American Statistical Association* 2000; **95**:596–607.
8. Kahn J, Lagakos S, Wulfson M, Cherng D, Miller M, Cherrington J, Hardy D, Beall G, Cooper R, Murphy R, Basgoz N, Ng E, Deeks S, Winslow D, Toole JJ, Coakley D. Efficacy and safety of Adefovir Dipivoxil with antiretroviral therapy: a randomized controlled trial. *Journal of the American Medical Association* 1999; **282**(24):2305–2312.
9. D'Aquila RT, Hughes MD, Johnson VA, Fischl MA, Sommadossi J, Song-heng L, Timpone J, Myers M, Basgoz N, Niu, M, Hirsch MS and the NIAID AIDS Clinical Trial Group Protocol 241 Investigators. Nevirapine, Zidovudine, and Didanosine compared with Zidovudine and Didanosine in patients with HIV-1 infection. *Annal of Internal Medicine* 1996; **124**:1019–1030.
10. Randles RH, Wolfe DA. *Introduction to the Theory of Nonparametric Statistics*. Wiley: New York, 1979.
11. Paxton WB, Coombs RW, McElrath MJ, Keefer MC, Hughes J, Sinangil F, Chernoff D, Demeter L, Williams B, Corey L. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with $\geqslant$400 CD4 lymphocytes: implications for applying measurements to individuals patients. *Journal of Infectious Diseases* 1997; **175**:247–254.