STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1646–1666
Published online 20 September 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3046

Incorporating intermediate binary responses into interim analyses of clinical trials: A comparison of four methods

Anne Whitehead[1,*,†], Marina Roshini Sooriyarachchi[2], John Whitehead[1] and Kim Bolland[1]

[1]*Medical and Pharmaceutical Statistics Research Unit, The University of Reading, Harry Pitt Building, Earley Gate, Whiteknights Road, Reading RG6 6FN, U.K.*
[2]*Department of Statistics, University of Colombo, Sri Lanka*

SUMMARY

In clinical trials with a long period of time between randomization and the primary assessment of the patient, the same assessments are often undertaken at intermediate times. When an interim analysis is conducted, in addition to the patients who have completed the primary assessment, there will be those who have till then undergone only intermediate assessments. The efficiency of the interim analysis can be increased by the inclusion of data from these additional patients. This paper compares four methods of increasing information based on model-free estimates of transition probabilities to incorporate intermediate assessments from patients who have not completed the trial. It is assumed that the observations are binary and that there is one intermediate assessment. The methods are the score and Wald approaches, each with the log-odds ratio and probability difference parameterizations. Simulations show that all four approaches have good properties in moderate to large sample sizes. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: binary data; longitudinal data; score test; sequential clinical trial; Wald test

1. INTRODUCTION

In many clinical trials there is a long period of time between randomization and the primary assessment of the patient. Examples include trials of severe head injury, which often use the Glasgow Outcome Scale at six months post-treatment, and stroke trials, which usually consider the Barthel Index, Modified Rankin Scale or NIH Stroke Scale at three months [1–4]. Despite the delay in obtaining assessments, a trial design incorporating interim analyses of the accumulating data

*Correspondence to: Anne Whitehead, Medical and Pharmaceutical Statistics Research Unit, University of Reading, Harry Pitt Building, Earley Gate, Whiteknights Road, Reading RG6 6FN, U.K.
†E-mail: p.a.whitehead@reading.ac.uk

Contract/grant sponsor: Novartis Pharma AG

Received 12 October 2006
Accepted 2 July 2007

can be advantageous, as it may reduce the time taken to reach a conclusion regarding the efficacy of a new treatment relative to a fixed sample size design. Indeed, sequential designs have been used for stroke trials, for example the ASCLEPIOS [5], RANTTAS [6] and eliprodil [unpublished] studies. When there is a long period of time between randomization and the primary assessment, the same assessment is often made at one or more intermediate times. When an interim analysis is conducted, in addition to the patients who have completed the primary assessment there will be those who have undergone only intermediate assessments. The efficiency of the interim analysis can be increased by the inclusion of these additional patients.

Although primary assessments in stroke and head injury trials are ordinal, they are often dichotomized. The focus of this paper is on methods for binary outcomes. In subsequent work, we hope to extend this approach to ordered categorical outcomes, which we realize would be a more acceptable approach to such trials in practice. Sooriyarachchi *et al*. [7] present a method for incorporating subjects with binary assessments taken at three fixed time points on each subject, with the third assessment time being the primary one. The primary assessment is predicted for those subjects with only one or two intermediate assessments using information about the transitions between successive assessments observed in the trial. There is no model fitted to the successive assessments; as the primary efficacy analysis concerns the relative treatment effect at the third assessment time, there is less interest in the earlier time points. Sooriyarachchi *et al*. derive a score test for the log-odds ratio based on the real and predicted values of the primary assessment. Marschner and Becker [8] consider a similar approach to the problem, but with only two time points. They present the Wald test based on the probability difference parameterization.

In this paper, the results of simulations comparing the properties of four methods for the case of binary observations taken at two time points are presented. The methods are the score and Wald approaches, each with the log-odds ratio and probability difference parameterizations, and so they include the method of [8] and a two-time point version of [7]. The test statistics for the four methods are presented in Section 2, and their use in a sequential design is described in Section 3. The methods are applied to data from a sequential trial in stroke in Section 4. The simulation study and its results are presented in Section 5 and, finally, Section 6 contains some concluding remarks.

2. TEST STATISTICS REPRESENTING TREATMENT EFFECT AND INFORMATION

In this paper, the primary response is the binary outcome at the second assessment time, and the measure of treatment difference is either the log-odds ratio (θ) or the probability difference (ψ). In this section, we first present the score and Wald statistics for these two parameters, based only on binary data from the second assessment. Then we consider the case in which subjects providing data only from the first assessment are additionally included. We show how the test statistics are modified to incorporate the predicted outcomes at the second assessment from these additional patients. Finally, we consider the handling of zero counts in the calculations.

2.1. Test statistics for binary outcomes at the second assessment

For a single binary outcome, the log-odds ratio θ is given by

$$\theta = \log\{p_1/(1-p_1)\} - \log\{p_2/(1-p_2)\} \tag{1}$$

DOI: 10.1002/sim

where p_1 and p_2 represent the success probabilities on the experimental treatment (T_1) and control treatment (T_2), respectively. The score approach consists of calculating the efficient score statistic (denoted by Z) and Fisher's observed information (denoted by V) for θ. These statistics (as given in Section 3.8.3 of [9]) are

$$Z^{(\theta)} = (n_2 S_1 - n_1 S_2)/n \quad \text{and} \quad V^{(\theta)} = (n_1 n_2 SF)/n^3 \tag{2}$$

where n_i, S_i and F_i are the number of patients, successes and failures on treatment T_i, $i = 1, 2$, and $n = n_1 + n_2$, $S = S_1 + S_2$ and $F = F_1 + F_2$. The use of score statistics in sequential clinical trials has been discussed by Whitehead [9].

For the probability difference parameter

$$\psi = p_1 - p_2 \tag{3}$$

the forms of the test statistics change to

$$\begin{aligned} Z^{(\psi)} &= n(n_2 S_1 - n_1 S_2)/(SF) \quad \text{and} \\ V^{(\psi)} &= n\{F^4 S_1 S_2 + S^2 F^2 (S_1 F_2 + S_2 F_1) + S^4 F_1 F_2\}/(SF)^3 \end{aligned} \tag{4}$$

as can be deduced directly, or from the expressions given in Section 3.6 of [10].

As an alternative to the score statistics, Wald statistics can be used. We set $Z = \hat{\theta}\{\text{se}(\hat{\theta})\}^{-2}$ and $V = \{\text{se}(\hat{\theta})\}^{-2}$, where $\hat{\theta}$ denotes the maximum likelihood estimate of θ and $\text{se}(\hat{\theta})$ its standard error. The maximum likelihood estimate of θ and its standard error are

$$\hat{\theta} = \log(S_1/F_1) - \log(S_2/F_2) \quad \text{and} \quad \text{se}(\hat{\theta}) = (S_1^{-1} + S_2^{-1} + F_1^{-1} + F_2^{-1})^{1/2} \tag{5}$$

(see, for example, Section 2.3.4 of [11]), from which test statistics, denoted by $Z_W^{(\theta)}$ and $V_W^{(\theta)}$, can be found. Use of such statistics in a sequential Wald test has been described by Cox [12]. The maximum likelihood estimate of the probability difference ψ and its standard error are

$$\hat{\psi} = (S_1/n_1) - (S_2/n_2) \quad \text{and} \quad \text{se}(\hat{\psi}) = \{(S_1 F_1)/n_1^3 + (S_2 F_2)/n_2^3\}^{1/2} \tag{6}$$

(see, for example, Section 2.3 of [11]), from which the statistics $Z_W^{(\psi)}$ and $V_W^{(\psi)}$ for a sequential Wald test can be deduced.

In each of the four methods, Z is an unstandardized statistic, constructed so that in large samples Z follows the normal distribution with mean θV (or ψV) and variance V. The statistic V is approximately proportional to the sample size and is a measure of the information in the trial.

2.2. Inclusion of patients with predicted second assessment outcomes

The test statistics will now be derived using data available from the intermediate assessment for patients who have not yet had their final assessment. Using the notation introduced in [7], the following data structure is assumed. Patients are randomized between an experimental treatment T_1 and a control treatment T_2 serially over time, and at time t_1 after their recruitment their condition is assessed and classified as being in either category C_1 (satisfactory) or category C_2 (unsatisfactory). Patients then continue in the trial till time t_2, when their condition is again assessed as being in either category C_1 or C_2. The primary response of interest is whether a patient is in category C_1 at time t_2, and this will be taken to represent the success of the treatment. Let $n_{ij,g}$ be the number

DOI: 10.1002/sim

of subjects randomized to treatment T_g in category C_i at time t_1 and category C_j at time t_2, and $n_{i*,g}$ be the number of subjects randomized to treatment T_g in category C_i at time t_1 and no time t_2 assessment, for $i,\ j,\ g=1,2$. We let $\circ$ denote summation over the subscripts 1 and 2, and $\bullet$ denote summation over the subscripts 1, 2 and $*$ (missing).

The probability that a patient on treatment T_g is in category C_i at t_1 and category C_j at t_2 will be denoted by $p_{ij,g}$. The conditional probabilities $q^{(1)}_{i,g}=P$ (category i at t_1; T_g) and $q^{(2)}_{ij,g}=P$(categories i at t_1 and j at $t_2|$ category i at t_1; T_g) can also be defined, and it follows that

$$p_{ij,g}=q^{(2)}_{ij,g}q^{(1)}_{i,g} \tag{7}$$

for $i,\ j,\ g=1,2$. Note that the probability that a patient changes category between the two assessments is allowed to differ between the two treatment groups. The term $N_{ij,g}$ will denote the number of patients on T_g who will eventually have outcome $(i,\ j)$ when all assessments have been completed. If some patients do not complete the second assessment, the values $N_{ij,g}$ will remain as latent observations. Depending on the pattern of incomplete and complete data available at the time that an analysis is conducted, the predicted value of $N_{ij,g}$, $e_{ij,g}$, is given by

$$e_{ij,g}=n_{ij,g}+n_{i*,g}q^{(2)}_{ij,g} \tag{8}$$

and the covariance of $N_{ij,g}$ and $N_{ij',g}$, $c_{i,(j),(j'),g}$, is given by

$$c_{i,(j),(j'),g}=n_{i*,g}q^{(2)}_{ij,g}\delta_{jj'}-n_{i*,g}q^{(2)}_{ij,g}q^{(2)}_{ij',g} \tag{9}$$

where $\delta_{jj'}=1$ if $j=j'$ and 0 otherwise.

To derive the test statistics, it is convenient to introduce the 'backwards' conditional probabilities $r^{(1)}_{j,g}=P$(category j at t_2; T_g) and $r^{(2)}_{ij,g}=P$(categories i at t_1 and j at $t_2|$ category j at t_2; T_g). It follows that

$$p_{ij,g}=r^{(2)}_{ij,g}r^{(1)}_{j,g} \tag{10}$$

for $i,\ j,\ g=1,2$.

Maximum likelihood estimates of the q's, e's and r's are given by

$$\hat{q}^{(1)}_{i,g}=n_{i\bullet,g}/n_{\circ\bullet,g},\quad \hat{q}^{(2)}_{ij,g}=n_{ij,g}/n_{i\circ,g} \tag{11}$$

$$\hat{e}_{ij,g}=n_{ij,g}+n_{i*,g}\hat{q}^{(2)}_{ij,g} \tag{12}$$

$$\hat{r}^{(1)}_{1,g}=\hat{e}_{\circ 1,g}/\hat{e}_{\circ\circ,g}\quad\text{and}\quad \hat{r}^{(2)}_{ij,g}=\hat{e}_{ij,g}/\hat{e}_{\circ j,g} \tag{13}$$

The four methods will now be developed in turn. In a rough order of ascending complexity the Wald statistics are considered first, and then the score statistics, with the probability difference version preceding the log-odds ratio in each case.

2.2.1. Wald statistics. For the probability difference parameter, ψ, where

$$\psi=r^{(1)}_{1,1}-r^{(1)}_{1,2} \tag{14}$$

DOI: 10.1002/sim

the maximum likelihood estimate, $\hat{\psi}$, is given by $\hat{\psi}=\hat{r}^{(1)}_{1,1}-\hat{r}^{(1)}_{1,2}$. Reparameterizing the log-likelihood ℓ ((A1) in Appendix) in terms of ψ, ζ, $r^{(2)}_{11,1}, r^{(2)}_{12,1}, r^{(2)}_{11,2}, r^{(2)}_{12,2}$, where $\zeta=r^{(1)}_{1,1}+r^{(1)}_{1,2}$, the Hessian matrix of second derivatives of ℓ with respect to ψ, ζ, $r^{(2)}_{11,1}, r^{(2)}_{12,1}, r^{(2)}_{11,2}, r^{(2)}_{12,2}$ in that order, $H(\psi_w)$, is obtained ((A2) in Appendix). The estimated variance of $\hat{\psi}$ is given by minus the leading element of the inverse of $H(\psi_w)$, replacing the r's, e's and q's by their maximum likelihood estimates given in (11)–(13). This estimated variance together with $\hat{\psi}$ can be used to calculate the statistics $Z^{(\psi)}_W$ and $V^{(\psi)}_W$ for a sequential Wald test. Marschner and Becker [8] presented their method in terms of parameters $r^{(1)}_{1,1}, r^{(1)}_{1,2}, q^{(1)}_{1,1}, q^{(1)}_{1,2}, q^{(2)}_{11,1}, q^{(2)}_{11,2}$, which should lead to the same estimates of ψ and its variance.

For the log-odds ratio parameter, θ, where

$$\theta=\log\{r^{(1)}_{1,1}/(1-r^{(1)}_{1,1})\}-\log\{r^{(1)}_{1,2}/(1-r^{(1)}_{1,2})\} \tag{15}$$

the maximum likelihood estimate, $\hat{\theta}$, is obtained by replacing $r^{(1)}_{1,g}$ with $\hat{r}^{(1)}_{1,g}$ in equation (15). The loglikelihood ℓ can be reparameterized in terms of θ, ϕ, $r^{(2)}_{11,1}, r^{(2)}_{12,1}, r^{(2)}_{11,2}, r^{(2)}_{12,2}$, where ϕ is the addition of the two log-odds terms in (15) instead of the subtraction. The Hessian matrix of second derivatives of ℓ with respect to θ, ϕ, $r^{(2)}_{11,1}, r^{(2)}_{12,1}, r^{(2)}_{11,2}, r^{(2)}_{12,2}$ in that order is given by $H(\theta_w)=H(\psi_w)$, with $w_m=r^{(1)}_{1,m}r^{(1)}_{2,m}/2$, $m=1,2$.

The estimated variance of $\hat{\theta}$ is given by minus the leading element of the inverse of $H(\theta_w)$, replacing the r's, e's and q's by their maximum likelihood estimates given in (11)–(13). This estimated variance together with $\hat{\theta}$ can be used to calculate the statistics $Z^{(\theta)}_W$ and $V^{(\theta)}_W$ for a sequential Wald test.

2.2.2. Score statistics. In order to derive the score statistics for ψ, restricted maximum likelihood estimates under the null hypothesis $H_0:\psi=0$ must be found. Setting $\partial\ell/\partial\zeta=0$ and $\psi=0$, the common restricted maximum likelihood estimate of $r^{(1)}_{1,g}$, $g=1,2$ is given by

$$\tilde{r}^{(1)}_{1,1}=\tilde{r}^{(1)}_{1,2}=\tilde{e}_{\circ 1,\circ}/\tilde{e}_{\circ\circ,\circ} \tag{16}$$

where $\tilde{e}_{\circ 1,\circ}=\tilde{e}_{\circ 1,1}+\tilde{e}_{\circ 1,2}$, and so on. Note that $\tilde{e}_{\circ\circ,\circ}=n_{\circ\bullet,\circ}$. The right-hand side of (16) is the null expected number of successes, divided by the total number of patients. It follows that the score statistic is

$$Z^{(\psi)}=\{n_{\circ\bullet,\circ}(n_{\circ\bullet,2}\tilde{e}_{\circ 1,1}-n_{\circ\bullet,1}\tilde{e}_{\circ 1,2})/(\tilde{e}_{\circ 1,\circ}\tilde{e}_{\circ 2,\circ})\}$$

In order to compute Z, restricted maximum likelihood estimates of all parameters are required, and these must satisfy $\partial\ell/\partial r^{(2)}_{1k,g}=0$, so that

$$\tilde{r}^{(2)}_{1j,g}=\tilde{e}_{1j,g}/\tilde{e}_{\circ j,g} \tag{17}$$

for g, $j=1,2$. The following iterative scheme is used: (i) the q's are estimated using their unrestricted maximum likelihood estimates given by (11), (ii) the e's are deduced from equation (12), (iii) the r's are then found using the restricted maximum likelihood equations (16) and (17), (iv)

DOI: 10.1002/sim

the e's are then found from the r's using equations (10), (7) and (12) in turn, and steps (iii) and (iv) are iterated to a solution.

The Hessian matrix $H(\psi_S)$ has the same structure as for the unrestricted case ((A2) in Appendix). Fisher's information, $V^{(\psi)}$, is given by minus the reciprocal of the leading element of the inverse of $H(\psi_S)$, replacing the r's, e's and q's by their restricted maximum likelihood estimates.

In order to derive the score statistics for θ, restricted maximum likelihood estimates under the null hypothesis $H_0: \theta = 0$ must be found. Setting $\partial\ell/\partial\phi = 0$ and $\theta = 0$, the common restricted maximum likelihood estimate of $r^{(1)}_{1,g}$, $g = 1, 2$ is given by (16). It follows that the score statistic is

$$Z^{(\theta)} = (n_{\circ\bullet,2}\tilde{e}_{\circ 1,1} - n_{\circ\bullet,1}\tilde{e}_{\circ 1,2})/n_{\circ\bullet,\circ}$$

The procedure for calculating $Z^{(\theta)}$ follows that for the calculation of $Z^{(\psi)}$.

The Hessian matrix $H(\theta_S) = H(\psi_S) + Y$, where Y is a 6×6 matrix, with entry y in positions (1,2) and (2,1), where y is given by

$$y = \{(\tilde{e}_{\circ 2,\circ} - \tilde{e}_{\circ 1,\circ})\tilde{e}_{\circ 1,\circ}\tilde{e}_{\circ 2,\circ}/4n^2_{\circ\bullet,\circ}\} \sum_{g=1}^{2}\sum_{j=1}^{2}(-1)^{g-1}(-1)^{j-1}\{\tilde{e}_{\circ j,g}/\tilde{e}_{\circ j,\circ}\}$$

and $w_1 = w_2 = \tilde{e}_{\circ 1,\circ}\tilde{e}_{\circ 2,\circ}/(2n^2_{\circ\bullet,\circ})$.

Fisher's information $V^{(\theta)}$ is given by minus the reciprocal of the leading element of the inverse of $H(\theta_S)$, replacing the r's, e's and q's by their restricted maximum likelihood estimates.

2.2.3. Handling of zero counts in the computations. The handling of zero counts is an important consideration in the calculation of the test statistics, although Marschner and Becker [8] omitted to provide any advice on this. If the denominator in any of the expressions (11)–(13) is zero, then the numerator will be zero too, and the ratio should be set to zero. If both $\hat{q}^{(1)}_{1,g}$ and $\hat{q}^{(1)}_{2,g}$ are set to zero, there are no patients at all on treatment T_g; hence, all four methods are bound to fail.

With regard to the Wald statistics, if $\hat{r}^{(2)}_{ij,g} = 0$, for any $i, j, g = 1, 2$, then the row and column relating to $r^{(2)}_{1j,g}$ in $H(\psi_w)$ and $H(\theta_w)$ should be removed before computation. If $\hat{r}^{(1)}_{j,g} = 0$, for any $j, g = 1, 2$, then the sequential Wald test cannot be implemented. The Wald test can only be used provided there are both successes and failures in both treatment groups at the second time point.

With regard to the score statistics, once a q term has been set to 0 in the first step of the iterative scheme, it will remain equal to 0 in subsequent steps. If $\tilde{e}_{\circ j,g} = 0$, then the algorithm forces $\tilde{r}^{(2)}_{1j,g}$ and $\tilde{r}^{(2)}_{2j,g}$ to be 0, so that the final estimate $\tilde{e}_{\circ j,g} = 0$, and the score statistic can still be calculated. The score statistic will become non-determinate only if there are no patients in the analysis. If $\tilde{r}^{(2)}_{ij,g} = 0$, for any $i, j, g = 1, 2$, then the row and column relating to $r^{(2)}_{1j,g}$ in $H(\psi_S)$ and $H(\theta_S)$ should be removed before computation. The score test can be used provided there are some successes and failures at the second time point irrespective of treatment group, i.e. $n_{\circ j,\circ} > 0$ for $j = 1, 2$.

3. SEQUENTIAL DESIGNS

Two sequential designs based on a boundaries approach are used to illustrate the methodology and for the simulations used to evaluate its properties. These are the triangular test [9] and the

DOI: 10.1002/sim

O'Brien and Fleming design [13], both of which are commonly used in practice. For each of these, the design specification is made in terms of the scalar parameter measuring the advantage of the experimental treatment over the control. Here, the procedures are described in terms of the log-odds ratio θ, although they would apply equally to the probability difference ψ. Selection of the type I and II error rates associated with a test of $H_0: \theta = 0$ *vs* $H_1: \theta \neq 0$ for a specific alternative value θ_R provides the stopping boundaries for the sequential clinical trial. These stopping boundaries may be presented as a plot of Z against V, for any of the pairs of statistics presented in Section 2. We let $V_{\max}$ denote the maximum value of V which lies on the stopping boundary.

The sequential procedure consists of a series of interim analyses, at the kth of which the current values Z_k and V_k of Z and V are compared with the stopping bounds L_k and U_k deduced from the design specification and the values of $V_1, \ldots, V_k$. Figure 1 illustrates this for the triangular test with Christmas tree boundaries [9]. For the triangular test, the trial continues if $Z_k \in (L_k, U_k)$ and $V_k < V_{\max}$ and is stopped otherwise. The case $Z_k \geqslant U_k$ and $V_k < V_{\max}$ corresponds to significant evidence that the experimental treatment is superior, while the case $Z_k \leqslant L_k$ and $V_k < V_{\max}$ may correspond to significant evidence that the experimental treatment is inferior or to the conclusion that no significant difference has been found. For the O'Brien and Fleming design $L_k = -U_k$, and the trial continues if $V_k < V_{\max}$ and $Z_k \in (-U_k, U_k)$ and is stopped otherwise. The case of $Z_k \geqslant U_k$ with $V_k < V_{\max}$ corresponds to significant evidence that the experimental treatment is superior, while the case $Z_k \leqslant -U_k$ with $V_k < V_{\max}$ corresponds to significant evidence that the experimental

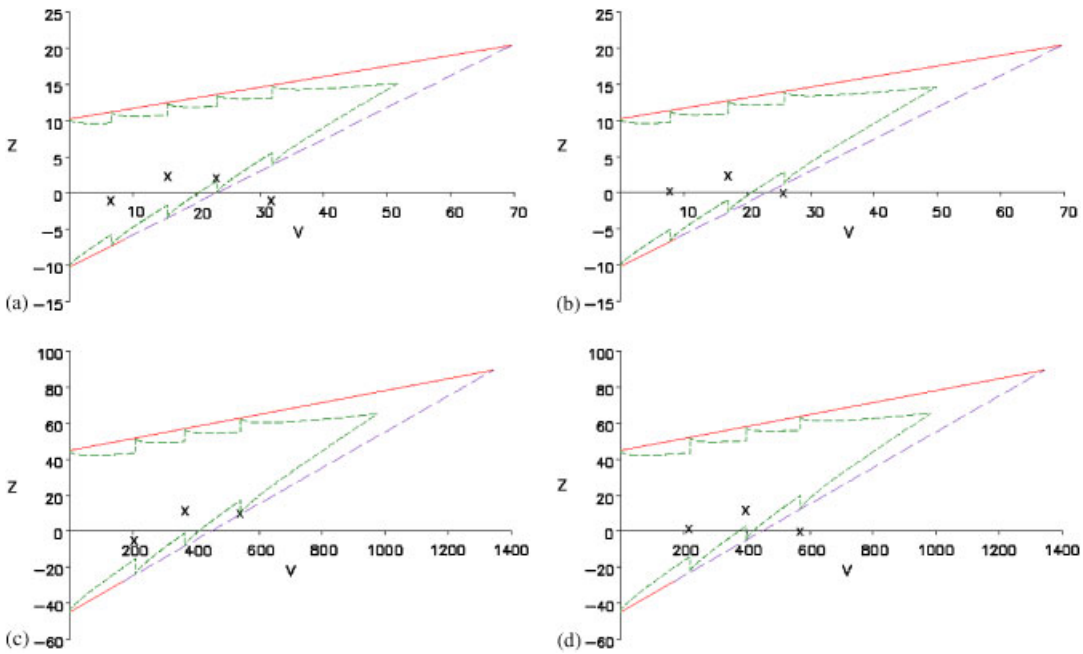

Figure 1. Group-sequential analysis using the triangular test of the binary outcome data from the stroke trial: (a) log-odds ratio—score test using only 90-day data; (b) log-odds ratio—score test using 30-day data as well; (c) probability difference—Wald test using only 90-day data; and (d) probability difference—Wald test using 30-day data as well.

DOI: 10.1002/sim

treatment is inferior. For either design, if the trial stops with $V_k \geqslant V_{\max}$, statistical significance cannot be determined by reference to the stopping boundaries. Once the trial is terminated a final analysis is undertaken, in which allowance is made for the interim analyses. A p-value and estimate and confidence interval for the treatment difference can be calculated using, for example, the methods described in Chapter 5 of Whitehead [9].

After a stopping boundary has been crossed, it is likely that data will continue to be collected. For example, a final analysis may be conducted when all randomized patients have completed the second assessment, and thus be based on the outcome at the second assessment. Such an analysis is referred to as an 'overrunning analysis'. There is a possibility that the study may stop because of a statistically significant result, which changes to a non-significant result at the final analysis. However, Sooriyarachchi *et al.* [14] have shown that this is likely to be a rare occurrence.

4. APPLICATION TO A SEQUENTIAL TRIAL IN STROKE

We illustrate the new methods using data from an international multi-centre phase III sequential clinical trial, comparing eliprodil with placebo for the treatment of patients following acute ischaemic stroke in the territory of the middle cerebral artery. Unfortunately, no medical manuscript on this trial has been published. The primary efficacy outcome was the Barthel Index at 90 days after randomization, grouped into the six ordered categories: 100 (complete recovery), 85–95, 60–80, 5–55, 0 (vegetative) and -5 (dead). Additional assessments were made at days 15, 30 and 60.

The study design used the triangular test, with a power of 0.9 to detect a log-odds ratio of 0.396, based on a proportional odds assumption, using a two-sided 5 per cent significance level. Interim analyses were planned after the 90-day assessment was available from 150 patients, and subsequently after every additional 150 patients. The 'last observation carried forward' principle was used to replace 90-day assessments recorded as missing. The lower boundary of the triangular test was crossed at the third interim analysis, with 90-day scores from 483 patients, indicating that the study should be stopped for futility. A final 'overrunning' analysis was conducted after the 875 randomized patients had provided the 90-day assessment. The overrunning analysis was an intention-to-treat analysis, which adjusted for the three interim analyses performed and gave an estimate for the log-odds ratio of -0.037 with 95 per cent confidence interval $(-0.307, 0.264)$ and $p = 0.796$.

For this paper, we selected the 30-day assessment as the intermediate outcome and considered the 844 patients out of the 875 randomized who had complete data for both the 30- and 90-day assessment times. The Barthel Index was dichotomized: $\geqslant 95$ (success), <95 (failure). Of the 431 patients in the eliprodil group, 92 (21.3 per cent) and 136 (31.6 per cent) had a successful outcome at days 30 and 90, respectively, and 91 (21.1 per cent) at both. Of the 413 subjects in the placebo group, 109 (26.4 per cent) and 141 (34.1 per cent) had a successful outcome at days 30 and 90, respectively, and 106 (25.7 per cent) at both.

For our illustrations, we set the same design objectives for both the log-odds ratio and the probability difference parameterizations. To ensure sufficient data for a triangular test boundary to be crossed, the clinically important treatment difference was chosen to correspond to success rates of 30 and 41 per cent in the placebo and eliprodil groups, respectively. A power of 90 per cent was therefore set to detect a log-odds ratio of 0.483 or a probability difference of 0.11. An equivalent fixed sample size design would require approximately 800 patients. In the reconstructed trials,

DOI: 10.1002/sim

Table I. Data from the stroke trial.

	12 July 1994			7 February 1995			29 June 1995			17 November 1995		
	Elip	Plac	Total	Elip	Plac	Total	Elip	Plac	Total	Elip	Plac	Total
(1,1)	9	16	25	31	32	63	48	45	93	58	67	125
(2,1)	7	3	10	15	9	24	23	16	39	33	26	59
(1,2)	1	0	1	1	1	2	1	1	2	1	2	3
(2,2)	56	58	114	104	107	211	163	153	316	207	206	413
(1,*)	7	3	10	6	6	12	6	14	20	16	14	30
(2,*)	12	9	21	19	18	37	15	19	34	21	21	42
(*,*)	9	8	17	12	13	25	18	14	32	23	17	40
Total randomized	101	97	198	188	186	374	274	262	536	359	353	712
At overrunning analysis												
(1,1)							55	60	115	80	87	167
(2,1)							30	23	53	38	29	67
(1,2)							1	2	3	1	2	3
(2,2)							188	177	365	240	235	475

Notes: Counts in the row labelled (i, j) give the numbers of subjects with assessments in Category C_i at 30 days and in Category C_j at 90 days, for $i, j = 1$ (success), 2 (failure) or * (missing).

patients were ordered according to their date of randomization, and interim analyses followed the original planned schedule.

The dichotomized assessments available at 30 and 90 days at each of the four interim analyses are shown in Table I. The corresponding values of Z and V computed according to the four methods described in Section 2 are displayed in Table II, using the 90-day assessments only, and then including the 30-day assessments. The final four rows in Table I and the final two columns in Table II relate to overrunning analyses. If at an interim analysis a stopping boundary is crossed, it is imagined that recruitment is stopped on the date of that analysis. The overrunning analysis is conducted when the 90-day data from all patients who had been randomized at the time of that interim analysis have been collected, and is based only on 90-day data. Figure 1(a) shows the results from the interim analyses using log-odds ratio score statistics based only on data from the 90-day assessment. The lower boundary of the triangular test is crossed at the fourth interim inspection, when 600 patients provide 90-day data. At this time, 72 patients would have provided only 30-day data and 40 would have been randomized but not assessed at all. The crossing of the lower stopping boundary indicates no significant difference between eliprodil and placebo; indeed there is a slight negative effect. An overrunning analysis conducted using 90-day assessments from all 712 randomized patients gives an estimate for the log-odds ratio of 0.097 with 95 per cent confidence interval $(-0.256, 0.493)$ and $p = 0.603$.

Reanalyses are now presented based on the same data, but include 30-day assessments for those patients who had completed the 30-day but not the 90-day assessment. The lower boundary of the triangular test is crossed at the third interim inspection, when the 90-day assessment for 450 patients and the 30-day assessment for 54 additional patients have been included (Figure 1(b)). A further 32 patients had been randomized but not yet assessed. An overrunning analysis using 90-day assessments from all 536 randomized patients gives an estimate for the log-odds ratio

DOI: 10.1002/sim

Table II. Test statistics for the stroke trial.

	90-day assessments only		90-day + 30-day assessments		For overrunning analysis	
Date	Z	V	Z	V	Z	V
Log-odds ratio score statistics						
12 July 1994	−1.033	6.704	0.197	7.910		
7 February 1995	2.210	15.442	2.413	17.071		
29 June 1995	2.067	23.274	−0.109	25.860	−0.881	28.821
17 November 1995	−0.693	31.893	0.322	35.474	0.014	39.271
Log-odds ratio Wald statistics						
12 July 1994	−1.030	6.670	0.197	7.919		
7 February 1995	2.207	15.405	2.410	17.038		
29 June 1995	2.064	23.219	−0.109	25.861	−0.881	28.826
17 November 1995	−0.693	31.891	0.322	35.474	0.014	39.271
Probability difference score statistics						
12 July 1994	−5.776	208.665	1.038	219.895		
7 February 1995	10.733	363.763	11.640	396.955		
29 June 1995	9.970	540.583	−0.511	569.307	−4.092	622.524
17 November 1995	−3.261	705.437	1.466	736.558	0.064	806.687
Probability difference Wald statistics						
12 July 1994	−5.795	210.163	1.038	219.877		
7 February 1995	10.749	364.769	11.650	397.615		
29 June 1995	9.990	542.757	−0.511	569.275	−4.091	622.258
17 November 1995	−3.261	705.506	1.466	736.550	0.064	806.688

of 0.016 with 95 per cent confidence interval $(-0.368, 0.434)$ and $p = 0.935$. Inclusion of the 30-day assessment has led to the same conclusion, with a 25 per cent reduction in sample size. The calculations for the log-odds ratio parameterization using Wald statistics are very similar.

Figure 1(c) shows the results from the interim analyses using probability difference Wald statistics based only on the data from the 90-day assessment. Although the design specification is the same as that for the log-odds ratio, the Z and V values are quite different from those in Figures 1(a) and (b). Using only the 90-day assessment the lower boundary is crossed at the third interim inspection, and an overrunning analysis using 90-day assessments from all 536 randomized patients gives an estimate for the probability difference of 0.0084, with 95 per cent confidence interval $(-0.0765, 0.1028)$ and $p = 0.852$. Inclusion of the patients with only 30-day data leads to an increase in V, although the boundary is still crossed at the third interim inspection (Figure 1(d)). An overrunning analysis shows that in this case addition of the incomplete data makes little difference: the estimate for the probability difference is 0.0132, with 95 per cent confidence interval $(-0.0732, 0.1077)$ and $p = 0.772$. The calculations for the probability difference parameterization using score statistics are very similar.

In the stroke trial, the number of patients providing data only for the 30-day assessment is small, ranging from 12 to 21 per cent of the number providing the 90-day assessment. For the log-odds ratio parameterization, this leads to an increase in V at an interim analysis of between 11 and 18 per cent. For the probability difference parameterization, the increase in V due to the 30-day assessments lies between 4 and 9 per cent. Estimates of the conditional probabilities

DOI: 10.1002/sim

$r^{(2)}_{11,1}, r^{(2)}_{22,1}, r^{(2)}_{11,2}, r^{(2)}_{22,2}$ from the 844 patients with both the 30- and 90-day data were 0.669, 0.997, 0.752 and 0.989, indicating that a failure at time 2 was almost certain to have been a failure at time 1, and that a success at time 2 would have had a high chance of being a success at time 1.

5. SIMULATION STUDIES

To explore the properties of the four repeated binary methods, two simulation studies were undertaken.

5.1. Performance in a single analysis

The first study compares the methods in terms of their performance in a single analysis. Two sample sizes were considered: 50 and 200 patients per treatment group. Three levels of completeness of data were considered; the proportion of patients providing data at time t_2 being set at 0.3, 0.5 and 0.6. For the control treatment, three probabilities of success at time t_2 were considered: $r^{(1)}_{1,2} = 0.1$, 0.5 and 0.8. For simulations under the null hypothesis of no treatment difference, the probabilities of success at time t_2 on the experimental treatment, $r^{(1)}_{1,1}$, were set to match the control values. For the case of 50 patients per group, the alternative hypotheses corresponding to $r^{(1)}_{1,2} = 0.1$, 0.5 and 0.8 were taken to be $r^{(1)}_{1,1} = 0.345$, 0.795 and 0.968, respectively. For the case of 200 patients per group, the corresponding alternative hypotheses were $r^{(1)}_{1,1} = 0.213$, 0.659 and 0.909, respectively. These alternatives were chosen to be those which would be detected with power 0.90 if complete data were available. This value should be borne in mind when evaluating the power achieved using the incomplete samples. Finally, three probabilities were set for achieving the same outcome at each of the two assessments, $r^{(2)}_{11,1} = r^{(2)}_{11,2} = r^{(2)}_{22,1} = r^{(2)}_{22,2} = 0.7$, 0.8 and 0.9. These values suggest that the patient's category is unlikely to change, as this is the situation most likely to yield an advantage for the incomplete data. The probabilities of not changing were set to be the same in the two treatment groups.

For each combination of settings, 10 000 simulations were performed. The Wald test can be calculated only when there are both successes and failures in both treatment groups at the second time point. The score test can be calculated provided there are some successes and failures at the second time point irrespective of the treatment group. Table III shows the proportion of times that the test statistic $Z/\sqrt{V}$ could be calculated and the null hypothesis rejected, that is $Z/\sqrt{V} \geqslant 1.96$, under the null hypothesis. The values in this table should be compared with 0.025. For the case of 50 patients per treatment group and a success rate of 0.1, the score test for the log-odds ratio for the repeated binary analysis inflates the type I error rate; the inflation factor increases as the per cent of the patients who provide data at the second time reduces. For all other scenarios the type I error rate for this test statistic is close to 0.025, with a tendency to be slightly higher. The score test for the probability difference tends to inflate the type I error rate by a larger amount and under more scenarios than that for the log-odds ratio. For the case of 50 patients per treatment group, the Wald test for the probability difference has a low type I error rate when the success rate is 0.1, but a slightly high type I error rate when the success rate is 0.5 or 0.8. For 200 patients per group, it gives results close to 0.025 and similar to those for the score test for the log-odds ratio. The occurrences of a low type I error rate for the Wald test are mainly due to the inability to calculate the Wald test due to zero successes or failures in one of the treatment groups at the

DOI: 10.1002/sim

Table III. Fixed sample size design: proportion of times H_0 is rejected to declare experimental treatment superior, under H_0 (nominal level = 0.025).

Number of subjects in each treatment group	$r_{1,2}^{(1)}$	$r_{1,1}^{(1)}$	$r_{i,g}^{(2)}$ $i,g=1,2$	Score test log-odds ratio			Wald test log-odds ratio			Score test probability difference			Wald test probability difference			Per cent of times the Wald test is not calculated		
				Per cent of subjects providing data at t_2			Per cent of subjects providing data at t_2			Per cent of subjects providing data at t_2			Per cent of subjects providing data at t_2			Per cent of subjects providing data at t_2		
				30	50	60	30	50	60	30	50	60	30	50	60	30	50	60
50	0.1	0.1	0.7	0.118	0.060	0.040	0.002	0.004	0.004	0.172	0.087	0.068	0.009	0.014	0.017	37.74	13.50	8.19
			0.8	0.132	0.062	0.045	0.003	0.005	0.007	0.178	0.089	0.070	0.014	0.016	0.018	37.16	13.77	8.80
			0.9	0.137	0.067	0.050	0.004	0.006	0.008	0.179	0.092	0.070	0.020	0.020	0.020	36.71	13.69	8.78
			t_2 only	0.013	0.027	0.028	0.001	0.003	0.004	0.177	0.090	0.072	0.002	0.010	0.015			
	0.5	0.5	0.7	0.034	0.030	0.025	0.029	0.028	0.023	0.032	0.030	0.025	0.038	0.033	0.028	0	0	0
			0.8	0.034	0.029	0.026	0.027	0.027	0.024	0.032	0.030	0.026	0.037	0.032	0.029	0.01	0	0
			0.9	0.033	0.028	0.024	0.027	0.025	0.022	0.032	0.028	0.024	0.034	0.030	0.026	0.01	0	0
			t_2 only	0.022	0.035	0.024	0.021	0.026	0.024	0.022	0.035	0.024	0.027	0.035	0.024			
	0.8	0.8	0.7	0.029	0.027	0.028	0.012	0.017	0.020	0.058	0.038	0.039	0.025	0.028	0.031	7.35	0.68	0.24
			0.8	0.033	0.028	0.026	0.014	0.017	0.018	0.061	0.037	0.034	0.029	0.029	0.029	7.34	0.77	0.26
			0.9	0.036	0.028	0.024	0.014	0.016	0.016	0.063	0.034	0.031	0.032	0.029	0.027	7.15	0.86	0.40
			t_2 only	0.023	0.027	0.025	0.003	0.016	0.017	0.060	0.043	0.038	0.013	0.026	0.031			
200	0.1	0.1	0.7	0.029	0.026	0.026	0.018	0.022	0.023	0.044	0.034	0.033	0.030	0.027	0.027	0.33	0	0
			0.8	0.032	0.027	0.027	0.019	0.023	0.024	0.041	0.033	0.034	0.033	0.028	0.027	0.30	0.01	0
			0.9	0.034	0.026	0.025	0.019	0.019	0.021	0.037	0.030	0.029	0.034	0.027	0.026	0.32	0.01	0
			t_2 only	0.029	0.027	0.027	0.017	0.021	0.024	0.049	0.034	0.034	0.034	0.027	0.028			
	0.5	0.5	0.7	0.027	0.029	0.028	0.027	0.028	0.027	0.027	0.029	0.028	0.028	0.030	0.029	0	0	0
			0.8	0.030	0.028	0.025	0.028	0.027	0.025	0.030	0.028	0.025	0.029	0.029	0.026	0	0	0
			0.9	0.031	0.025	0.024	0.029	0.023	0.024	0.031	0.025	0.024	0.030	0.025	0.024	0	0	0
			t_2 only	0.026	0.030	0.023	0.026	0.030	0.023	0.026	0.030	0.023	0.026	0.030	0.024			
	0.8	0.8	0.7	0.025	0.026	0.024	0.022	0.024	0.023	0.028	0.027	0.025	0.027	0.027	0.025	0	0	0
			0.8	0.029	0.025	0.024	0.024	0.024	0.023	0.030	0.028	0.026	0.029	0.026	0.025	0	0	0
			0.9	0.029	0.028	0.026	0.023	0.025	0.025	0.029	0.029	0.028	0.028	0.028	0.027	0	0	0
			t_2 only	0.024	0.026	0.023	0.023	0.023	0.022	0.030	0.029	0.025	0.026	0.026	0.024			

DOI: 10.1002/sim

second timepoint (Table III). The Wald test for the log-odds ratio has a low type I error rate for the case of 50 patients per treatment group and a success rate of 0.1 or 0.8.

Table IV shows the proportion rejecting the null hypothesis under the alternative hypothesis. For the smaller sample size of 50 patients per treatment, the score test for the probability difference tends to have the highest power out of all four tests, although for the larger sample size of 200 patients per treatment it is similar to the score test for the log-odds ratio. The Wald test for the log-odds ratio generally has the lowest power out of the four tests. The Wald test for the probability difference and the score test for the log-odds ratio give similar results, except when the success rate in the control group is 0.8 with 50 patients per treatment. Compared with the analyses of patients who have completed the second assessment, the repeated binary approach increases the power, by about 20 per cent for the scenarios in which there is a probability of 0.9 of the same outcome at each time point, and 50 or 60 per cent of the patients provide data at the second time point: when only 30 per cent of the patients provide data at the second time point, the power can increase by more than 50 per cent.

In summary, both power and type I error rates generally increase in the order: Wald test for the log-odds ratio, Wald test for the probability difference, score test for the log-odds ratio and score test for the probability difference. For sample sizes of 200 subjects per treatment arm, all tests with the exception of the score test for the probability difference have type I error rates close to the nominal level. For sample sizes of 50 subjects per treatment arm, the type I error rates are reasonably close to the nominal level for all tests when the success rate is 0.5, but deviate more from the nominal level as the success rate moves away from 0.5. When the success rate is 0.1, none of the methods is satisfactory. Generally, type I error rates deviate more from the nominal level as the percentage of subjects with data at the second time point reduces.

5.2. Performance within a sequential setting

The second simulation study compares the four methods within a sequential setting. Two sequential designs were selected, the triangular test and the O'Brien and Fleming design. A two-sided 5 per cent significance level was selected. For the control treatment, three probabilities of success at time t_2 were considered: $r_{1,2}^{(1)} = 0.1$, 0.5 and 0.8. Under the null hypothesis of no treatment difference the probabilities of success at time t_2 on the experimental treatment, $r_{1,1}^{(1)}$, were set to match the control values, and under the alternative hypothesis they were set to 0.177, 0.613 and 0.882, respectively. These alternatives were chosen to be those which would be detected with power 0.90 if complete data were available from a fixed sample size design with 400 patients per treatment group. This resulted in three scenarios for each method. Whereas the fixed sample size scenarios in Section 5.1 were chosen to investigate the methods with small sample sizes, the scenarios for the sequential setting have been chosen to represent the larger sample sizes which might occur in practice. The specifications for the stopping boundaries for the designs are presented in Table V.

The probabilities of the same outcome at each of the two assessments, $r_{11,1}^{(2)} = r_{11,2}^{(2)} = r_{22,1}^{(2)} = r_{22,2}^{(2)}$, were set to be 0.7, 0.8 and 0.9. Additionally, scenarios with $r_{11,1}^{(2)} = r_{22,1}^{(2)} = 0.7$ and $r_{11,2}^{(2)} = r_{22,2}^{(2)} = 0.9$ were used to investigate the situation in which the probability of the same outcome is different in the two treatment groups. These scenarios can be used to assess performance under the null hypothesis that there is no treatment difference at assessment 2, in the presence of a treatment difference at assessment 1.

DOI: 10.1002/sim

Table IV. Fixed sample size design: proportion of times H_0 is rejected to declare experimental treatment superior, under H_1.

Number of subjects in each treatment group	$r_{1,2}^{(1)}$	$r_{1,1}^{(1)}$	$r_{ii,g}^{(2)}$ $i,g=1,2$	Score test log-odds ratio: Per cent of subjects providing data at t_2			Wald test log-odds ratio: Per cent of subjects providing data at t_2			Score test probability difference: Per cent of subjects providing data at t_2			Wald test probability difference: Per cent of subjects providing data at t_2			Per cent of times the Wald test is not calculated: Per cent of subjects providing data at t_2		
				30	50	60	30	50	60	30	50	60	30	50	60	30	50	60
50	0.1	0.345	0.7	0.485	0.612	0.676	0.249	0.517	0.616	0.493	0.663	0.723	0.400	0.605	0.693	21.26	7.09	4.31
			0.8	0.539	0.656	0.712	0.297	0.553	0.649	0.530	0.692	0.751	0.457	0.640	0.723	21.30	7.10	4.49
			0.9	0.652	0.730	0.771	0.442	0.610	0.710	0.631	0.752	0.798	0.592	0.704	0.769	20.76	7.06	4.56
			t_2 only	0.379	0.592	0.642	0.144	0.502	0.583	0.493	0.677	0.734	0.311	0.626	0.676			
	0.5	0.795	0.7	0.476	0.652	0.717	0.405	0.627	0.699	0.478	0.679	0.740	0.473	0.668	0.732	3.33	0.28	0.09
			0.8	0.541	0.704	0.746	0.462	0.677	0.726	0.529	0.722	0.770	0.521	0.710	0.756	3.41	0.28	0.10
			0.9	0.669	0.786	0.813	0.608	0.756	0.797	0.654	0.799	0.830	0.640	0.784	0.818	3.32	0.39	0.13
			t_2 only	0.404	0.633	0.678	0.316	0.611	0.664	0.416	0.631	0.699	0.432	0.646	0.698			
	0.8	0.968	0.7	0.215	0.481	0.570	0.061	0.190	0.281	0.641	0.676	0.717	0.129	0.328	0.441	63.69	45.12	37.33
			0.8	0.287	0.529	0.602	0.092	0.219	0.302	0.647	0.701	0.726	0.179	0.385	0.477	63.67	45.79	38.35
			0.9	0.449	0.603	0.662	0.190	0.298	0.365	0.681	0.720	0.757	0.296	0.452	0.541	62.62	44.66	38.32
			t_2 only	0.238	0.506	0.589	0.013	0.184	0.280	0.646	0.693	0.731	0.050	0.289	0.427			
200	0.1	0.213	0.7	0.441	0.626	0.703	0.407	0.611	0.692	0.477	0.628	0.702	0.469	0.614	0.687	0.16	0	0
			0.8	0.488	0.660	0.726	0.445	0.642	0.714	0.504	0.656	0.719	0.511	0.644	0.710	0.16	0	0
			0.9	0.583	0.729	0.774	0.521	0.709	0.763	0.578	0.716	0.764	0.597	0.710	0.757	0.15	0.01	0
			t_2 only	0.401	0.612	0.691	0.387	0.594	0.679	0.454	0.654	0.737	0.451	0.590	0.676			
	0.5	0.659	0.7	0.479	0.665	0.734	0.472	0.661	0.731	0.482	0.663	0.731	0.483	0.668	0.735	0	0	0
			0.8	0.542	0.713	0.772	0.532	0.707	0.769	0.545	0.710	0.769	0.540	0.713	0.772	0	0	0
			0.9	0.683	0.790	0.830	0.667	0.786	0.828	0.683	0.787	0.828	0.669	0.788	0.830	0	0	0
			t_2 only	0.434	0.635	0.710	0.436	0.635	0.705	0.437	0.633	0.706	0.434	0.633	0.720			
	0.8	0.909	0.7	0.427	0.629	0.701	0.390	0.610	0.687	0.468	0.659	0.723	0.440	0.637	0.706	0.40	0.01	0
			0.8	0.461	0.662	0.719	0.413	0.642	0.704	0.484	0.681	0.740	0.467	0.668	0.723	0.45	0.01	0
			0.9	0.566	0.719	0.770	0.503	0.698	0.755	0.567	0.729	0.781	0.560	0.721	0.773	0.37	0	0
			t_2 only	0.390	0.613	0.688	0.368	0.588	0.677	0.449	0.640	0.715	0.424	0.613	0.696			

DOI: 10.1002/sim

Table V. Specification of the sequential designs used in the simulations.

			Upper boundary		Lower boundary		
	$r_{1,2}^{(1)}$	$r_{1,1}^{(1)}$	Intercept	Slope	Intercept	Slope	$V_{\max}$
Triangular test							
Log-odds ratio	0.1	0.177	7.438	0.2014	−7.438	0.6042	36.93
	0.5	0.613	10.690	0.1401	−10.690	0.4204	76.29
	0.8	0.882	7.867	0.1904	−7.867	0.5712	41.32
Probability difference	0.1	0.177	63.72	0.02351	−63.72	0.07052	2711
	0.5	0.613	43.52	0.03442	−43.52	0.10326	1264
	0.8	0.882	60.02	0.02496	−60.02	0.07487	2405
O'Brien and Fleming test							
Log-odds ratio	0.1	0.177	11.249	0	−11.249	0	25.19
	0.5	0.613	16.167	0	−16.167	0	52.03
	0.8	0.882	11.898	0	−11.898	0	28.18
Probability difference	0.1	0.177	96.38	0	−96.38	0	1849
	0.5	0.613	65.81	0	−65.81	0	862
	0.8	0.882	90.77	0	−90.77	0	1640

The recruitment rate to the study was fixed at 50 patients per month, and the two assessment time points were 1 and 3 months. Interim analyses were undertaken 6 months after the first patient was randomized, and then every 3 months until a boundary was crossed. Of the 300 (450) patients randomized at the first (second) interim analysis, 150 (300) would provide data from the second assessment, and similarly at later analyses. At each interim analysis 100 patients would provide data from the first assessment only and a further 50 would have been randomized, but not been assessed.

For each scenario, 10 000 simulations were performed. Table VI shows the proportion of times that an upper stopping boundary was crossed under the null hypothesis: the values in this table should be compared with 0.025. All four methods maintain a type I error rate of about 0.025. The better performance of the tests in these simulations relative to those in Section 5.1 is likely to be due both to the larger sample sizes and to the way in which sequential tests overcome some of the discreteness problems inherent in fixed sample tests concerning binary data which can yield only a limited set of test statistic values. Accurate type I errors appear to be maintained when there is a treatment difference at time t_1, but not at time t_2. Table VI also shows the average number of patients randomized at the point when a stopping boundary is crossed under the null hypothesis. It can be seen that when the success probability is 0.5, the average sample sizes are similar across all four methods. However, when the success probability is 0.1, sample sizes are larger for the log-odds ratio parameterization, and when it is 0.8 they are larger for the probability difference parameterization. This is because the sample size for the log-odds ratio is approximately equal to $4V/\{r(1-r)\}$, where $r=(r_{1,1}^{(1)}+r_{1,2}^{(1)})/2$, whereas the sample size for the probability difference is approximately equal to $4V\{r(1-r)\}$. If r under the null hypothesis moves further away from 0.5 than r under the alternative hypothesis, then sample sizes will be larger for the log-odds ratio parameterization. If r under the null hypothesis moves closer to 0.5 than r under the alternative hypothesis, then sample sizes will be larger for the probability difference parameterization. The triangular test is used if it is desirable to stop early under the null hypothesis for futility. Sample

DOI: 10.1002/sim

Table VI. Sequential designs: under H_0—proportion of times H_0 is rejected to declare experimental treatment superior (nominal level = 0.025) and average number of subjects randomized when stopping boundary crossed.

					Triangular test						O'Brien and Fleming design					
					Log-odds ratio			Probability difference			Log-odds ratio			Probability difference		
$r_{1,2}^{(1)}$	$r_{1,1}^{(1)}$	$r_{ii,g}^{(2)}$ $i,g=1,2$	$q_{1,2}^{(1)}$	$q_{1,1}^{(1)}$	Score test	Wald test		Score test	Wald test		Score test	Wald test		Score test	Wald test	
0.1	0.1	0.7	0.34	0.34	0.026	724	0.023	734	0.028	500	0.026	495	0.024	1329	0.025	1337	0.029	874	0.028	867
		0.8	0.26	0.26	0.025	715	0.022	728	0.032	496	0.026	491	0.028	1317	0.026	1326	0.028	866	0.026	860
		0.9	0.18	0.18	0.024	698	0.019	705	0.027	475	0.026	469	0.024	1297	0.020	1307	0.028	847	0.025	839
		t_2 only			0.027	729	0.025	739	0.027	507	0.025	502	0.023	1335	0.025	1342	0.028	879	0.028	873
0.5	0.5	0.7	0.5	0.5	0.024	576	0.025	578	0.024	589	0.025	588	0.026	1039	0.026	1039	0.023	1040	0.025	1040
		0.8	0.5	0.5	0.028	561	0.027	562	0.027	571	0.029	570	0.024	1038	0.023	1039	0.023	1038	0.025	1037
		0.9	0.5	0.5	0.025	531	0.024	532	0.024	543	0.025	542	0.029	1035	0.029	1035	0.026	1037	0.028	1036
		t_2 only			0.024	594	0.024	597	0.024	601	0.026	599	0.026	1039	0.026	1040	0.022	1041	0.022	1041
0.8	0.8	0.7	0.62	0.62	0.027	526	0.026	531	0.026	683	0.024	679	0.023	904	0.024	908	0.026	1250	0.025	1246
		0.8	0.68	0.68	0.025	510	0.022	515	0.024	670	0.025	668	0.026	897	0.022	900	0.026	1229	0.024	1228
		0.9	0.74	0.74	0.025	486	0.023	489	0.028	644	0.029	641	0.025	882	0.025	885	0.029	1202	0.025	1200
		t_2 only			0.026	535	0.025	539	0.026	691	0.024	689	0.025	911	0.023	916	0.028	1262	0.026	1258
0.1	0.1	0.7, 0.9*	0.18	0.34	0.023	710							0.024	1314						
0.8	0.8	0.7, 0.9*	0.74	0.62	0.025	508							0.024	895						

*$r_{ii,1}^{(2)}=0.7$, $i=1,2$; $r_{ii,2}^{(2)}=0.9$, $i=1,2$.

DOI: 10.1002/sim

Table VII. Sequential designs: under H_1—proportion of times H_0 is rejected to declare experimental treatment superior and average number of subjects randomized when stopping boundary crossed.

					Triangular test								O'Brien and Fleming design							
					Log-odds ratio				Probability difference				Log-odds ratio				Probability difference			
$r_{1,2}^{(1)}$	$r_{1,1}^{(1)}$	$r_{ii,g}^{(2)}$ $i, g = 1, 2$	$q_{1,2}^{(1)}$	$q_{1,1}^{(1)}$	Score test		Wald test		Score test		Wald test		Score test		Wald test		Score test		Wald test	
0.1	0.177	0.7	0.34	0.371	0.896	681	0.901	710	0.898	683	0.908	672	0.900	754	0.903	791	0.913	769	0.899	757
		0.8	0.26	0.306	0.896	672	0.895	700	0.897	669	0.907	658	0.902	744	0.907	781	0.919	754	0.907	741
		0.9	0.18	0.242	0.893	643	0.895	678	0.895	646	0.900	633	0.900	722	0.903	758	0.914	731	0.904	720
		t_2 only			0.896	689	0.902	718	0.900	690	0.906	679	0.898	761	0.901	797	0.917	777	0.898	766
0.5	0.613	0.7	0.5	0.545	0.899	667	0.898	672	0.898	669	0.897	661	0.905	738	0.904	744	0.907	738	0.908	730
		0.8	0.5	0.568	0.893	650	0.898	656	0.900	654	0.901	645	0.909	722	0.906	731	0.910	722	0.907	715
		0.9	0.5	0.591	0.895	622	0.896	625	0.898	623	0.901	616	0.921	697	0.921	704	0.922	697	0.922	689
		t_2 only			0.899	679	0.901	684	0.899	681	0.897	673	0.903	746	0.902	758	0.901	751	0.903	742
0.8	0.882	0.7	0.62	0.653	0.894	677	0.894	704	0.901	677	0.896	673	0.899	753	0.899	781	0.910	765	0.903	754
		0.8	0.68	0.729	0.892	666	0.896	691	0.903	668	0.896	662	0.900	740	0.904	770	0.913	750	0.908	737
		0.9	0.74	0.806	0.893	641	0.901	670	0.897	641	0.893	633	0.903	716	0.904	748	0.910	727	0.903	715
		t_2 only			0.895	686	0.893	712	0.898	685	0.893	680	0.899	764	0.902	788	0.910	773	0.904	764
0.1	0.1	0.7, 0.9*	0.18	0.371	0.895	666							0.904	738						
0.8	0.8	0.7, 0.9*	0.74	0.653	0.894	662							0.901	733						

*$r_{ii,1}^{(2)} = 0.7$, $i = 1, 2$; $r_{ii,2}^{(2)} = 0.9$, $i = 1, 2$.

DOI: 10.1002/sim

size reductions of up to about 10 per cent can be seen for the triangular test for scenarios in which there is a probability of 0.9 of the same outcome at each time point. The O'Brien and Fleming test is used if it is undesirable to stop early under the null hypothesis. Sample size reductions are smaller for the O'Brien and Fleming test, because under the null hypothesis the likely outcome is that the vertical boundary at $V = V_{\max}$ will be crossed.

Table VII shows the proportion of times that an upper stopping boundary was crossed under the alternative hypothesis: the values in this table should be compared with 0.9. All tests maintain power for all scenarios, including the cases in which the treatment difference at time t_1 is not the same as that at time t_2. Average sample sizes under the alternative hypothesis are also shown. These show comparable sample sizes across all four methods, as expected. Sample size reductions due to the incorporation of 30-day data of up to about 7 per cent can be seen.

6. DISCUSSION

For interim analyses of clinical trials with long patient follow-up before the primary assessment, inclusion of intermediate assessments from patients still being followed up makes efficient use of the available data. In this paper, we have investigated the properties of four methods for the case of binary observations taken at two time points. All four methods estimate transition probabilities from the data collected in the trial, and none assume a model linking the two assessments.

Comparisons of the four methods in the analysis of a single dataset show that both power and type I error rates generally increase in the order: Wald test for the log-odds ratio, Wald test for the probability difference, score test for the log-odds ratio and score test for the probability difference. For sample sizes of 200 subjects per treatment arm, all tests with the exception of the score test for the probability difference have type I error rates close to the nominal level. However, for the smaller sample size of 50 subjects per treatment arm, the type I error rate deviates further from the nominal level as the success rate moves away from 0.5: the score tests tend to have an inflated type I error rate, whereas the Wald tests have too low a type I error rate. Generally, type I error rates deviate more from the nominal level as the percentage of subjects with data at the second time point reduces. One advantage of the score test over the Wald test is that it can be calculated provided that there are both successes and failures at the second time point, irrespective of the treatment group: the Wald test requires both successes and failures *in each treatment group* at the second time point. Consequently, when success rates are close to 0 or 1 in one or both of the treatment arms and the sample size is small, there will be more occasions when the Wald test cannot be calculated. In the sequential settings, all four tests demonstrated good properties, even with low success probabilities, although sample sizes were quite large.

In the stroke trial, the number of patients providing data for the 30-day assessment but not the 90-day assessment was small, lying between 12 and 21 per cent of the number providing the 90-day assessment. This led to an increase in V at an interim analysis of between 11 and 18 per cent for the log-odds ratio and between 4 and 9 per cent for the probability difference. In the scenarios investigated for the sequential settings, the savings in the sample size were modest. Under the null hypothesis, sample size reductions of up to about 10 per cent were seen for the triangular test. Sample size reductions were smaller for the O'Brien and Fleming test, because under the null hypothesis the likely outcome is that the vertical boundary is crossed. Under the alternative hypothesis there were comparable sample size reductions across all four methods, of up to about 7 per cent. The magnitude of the sample size reduction depends on the number of patients who

DOI: 10.1002/sim

have completed assessment 1 but not yet reached assessment 2. With expected sample sizes of 600, only 100 (17 per cent) of the patients would have been in this category. Thus, larger savings in sample size might be expected if this proportion is higher.

For a sequential clinical trial, the expected sample sizes under the null hypothesis are dependent on the choice of the log-odds ratio or the probability difference, and on the average success rate across the two treatment groups. If the average success rate under the null hypothesis moves further away from 0.5 than it does when under the alternative hypothesis, then sample sizes will be larger for the log-odds ratio parameterization. If the average success rate under the null hypothesis moves closer to 0.5 than it does under the alternative hypothesis, then sample sizes will be larger for the probability difference parameterization. On the other hand, the expected sample sizes under the alternative hypothesis are similar for the two parameterizations.

Programs written in C to calculate the Z and V statistics for the four methods for an individual time point are available from the authors. In the case where there are no zeros in the first four rows in Table III, the Wald statistics may also be obtained using SAS PROC NLMIXED, by specifying the log-likelihood in the form of equation (A1) in the Appendix, and re-expressing $r^{(1)}_{1,1}$ and $r^{(1)}_{1,2}$ in terms of ψ and ζ for the probability difference or θ and ϕ for the log-odds ratio.

The methods described in this paper can be extended to include adjustment for categorical covariates, using the approach of Chapter 7.2 of [9]. To achieve this, the statistics Z and V are computed separately for each stratum created by the combination of covariate values, and then the Z's are summed and the V's are summed to provide the overall Z and V values. Also, the extension to ordered categorical data could be considered, in which the treatment difference at the primary assessment was expressed as an odds ratio from a proportional odds model.

APPENDIX

In order to incorporate the intermediate assessment from patients who have not yet had their final assessment, we express the log-likelihood ℓ in terms of the 'backwards' conditional probabilities $r^{(1)}_{j,g}$ and $r^{(2)}_{ij,g}$ as follows:

$$\ell = \sum_{g=1}^{2}\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij,g}\log(r^{(2)}_{ij,g}r^{(1)}_{j,g}) + \sum_{g=1}^{2}\sum_{i=1}^{2} n_{i*,g}\log(r^{(2)}_{i1,g}r^{(1)}_{1,g} + r^{(2)}_{i2,g}r^{(1)}_{2,g}) \tag{A1}$$

If H is the Hessian matrix of second derivatives of ℓ with respect to the six r terms in the order $r^{(2)}_{11,1}, r^{(2)}_{12,1}, r^{(1)}_{1,1}, r^{(2)}_{11,2}, r^{(2)}_{12,2}, r^{(1)}_{1,2}$, then H is a 6×6 block diagonal matrix, where each block is a 3×3 matrix of the form

$$\begin{pmatrix} a_{1m} & b_m & c_{1m} \\ b_m & a_{2m} & c_{2m} \\ c_{1m} & c_{2m} & d_m \end{pmatrix}$$

where

$$a_{km} = -\sum_{i=1}^{2}\{e_{ik,m}/(r^{(2)}_{ik,m})^2\} + \sum_{i=1}^{2}\{c_{i,(k),(k),m}/(r^{(2)}_{ik,m})^2\}$$

DOI: 10.1002/sim

$$b_m = \sum_{i=1}^{2} \{c_{i,(1),(2),m}/(r_{i1,m}^{(2)} r_{i2,m}^{(2)})\}$$

$$c_{km} = \sum_{i=1}^{2} \sum_{j'=1}^{2} \{(-1)^{i-1}(-1)^{j'-1} c_{i,(k),(j'),m}/(r_{j',m}^{(1)} r_{ik,m}^{(2)})\}$$

and

$$d_m = -\sum_{j=1}^{2} \{e_{\circ j,m}/(r_{j,m}^{(1)})^2\} + \sum_{j=1}^{2} \sum_{j'=1}^{2} \{(-1)^{j-j'} c_{\circ,(j),(j'),m}/(r_{j,m}^{(1)} r_{j',m}^{(1)})\}$$

for $k, m = 1, 2$.

For the probability difference parameter ψ, where

$$\psi = r_{1,1}^{(1)} - r_{1,2}^{(1)}$$

the log-likelihood ℓ can be reparameterized in terms of ψ, ζ, $r_{11,1}^{(2)}, r_{12,1}^{(2)}, r_{11,2}^{(2)}, r_{12,2}^{(2)}$, where $\zeta = r_{1,1}^{(1)} + r_{1,2}^{(1)}$, and $H(\psi_w)$, the Hessian matrix of second derivatives of ℓ with respect to ψ, ζ, $r_{11,1}^{(2)}, r_{12,1}^{(2)}, r_{11,2}^{(2)}, r_{12,2}^{(2)}$ in that order, is given by

$$H(\psi_w) = \begin{pmatrix} w_1^2 d_1 + w_2^2 d_2 & w_1^2 d_1 - w_2^2 d_2 & w_1 c_{11} & w_1 c_{21} & -w_2 c_{12} & -w_2 c_{22} \\ w_1^2 d_1 - w_2^2 d_2 & w_1^2 d_1 + w_2^2 d_2 & w_1 c_{11} & w_1 c_{21} & w_2 c_{12} & w_2 c_{22} \\ w_1 c_{11} & w_1 c_{11} & a_{11} & b_1 & 0 & 0 \\ w_1 c_{21} & w_1 c_{21} & b_1 & a_{21} & 0 & 0 \\ -w_2 c_{12} & w_2 c_{12} & 0 & 0 & a_{12} & b_2 \\ -w_2 c_{22} & w_2 c_{22} & 0 & 0 & b_2 & a_{22} \end{pmatrix} \tag{A2}$$

with $w_m = 0.5$, $m = 1, 2$.

ACKNOWLEDGEMENTS

The authors wish to thank Novartis Pharma AG, Switzerland, for their financial support of this project.

REFERENCES

1. Pickard JD, Murray GD, Illingworth R, Shaw MDM, Teasdale GM, Foy PM, Humphrey PRD, Lang DA, Nelson R, Richards P, Sinar J, Bailey S, Skene A. Effect of oral nimodipine on cerebral infarction and outcome after subarachnoid haemorrhage: British aneurysm nimodipine trial. *British Medical Journal* 1989; **298**:636–642.
2. Marshall LF, Maas AI, Marshall SB, Bricolo A, Fearnside M, Iannotti F, Klauber MR, Lagarrigue J, Lobato R, Persson L, Pickard JD, Piek J, Servadei F, Wellis GN, Morris GF, Means ED, Musch B. A multicenter trial on the efficacy of using tirilizad mesylate in cases of head injury. *Journal of Neurosurgery* 1998; **89**:519–525.
3. NINDS, The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *The New England Journal of Medicine* 1995; **333**:1581–1587.
4. Lees KR, Asplund K, Carolei A, Davis S, Diener HC, Kaste M, Orgogozo JM, Whitehead J. for the GAIN International Investigators. Glycine antagonist (gavestinel) in neuroprotection (GAIN International) in patients with acute stroke: a randomised controlled trial. *The Lancet* 2000; **335**:1949–1954.

DOI: 10.1002/sim

5. Whitehead J. Application of sequential methods to a phase III clinical trial in stroke. *Drug Information Journal* 1993; **27**:733–740.
6. The RANTTAS Investigators. A randomized trial of tirilazad mesylate in patients with acute stroke (RANTTAS). *Stroke* 1996; **27**:1453–1458.
7. Sooriyarachchi MR, Whitehead J, Whitehead A, Bolland K. The sequential analysis of repeated binary responses: a score test for the case of three time points. *Statistics in Medicine* 2006; **25**:2196–2214.
8. Marschner IC, Becker SL. Interim monitoring of clinical trials based on long-term binary endpoints. *Statistics in Medicine* 2001; **20**:177–192.
9. Whitehead J. *The Design and Analysis of Sequential Clinical Trials* (revised 2nd edn). Wiley: Chichester, 1997.
10. Whitehead J. *The Design and Analysis of Sequential Clinical Trials* (1st edn). Ellis Horwood: Chichester, 1983.
11. Collett D. *Modelling Binary Data* (2nd edn). Chapman & Hall/CRC: London, Boca Raton, FL, 2003.
12. Cox DR. Large sample sequential tests for composite hypotheses. *Sankhyā* 1963; **25**:5–12.
13. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
14. Sooriyarachchi MR, Whitehead J, Matsushita T, Bolland K, Whitehead A. Incorporating data received after a sequential trial has stopped into the final analysis: implementation and comparison of methods. *Biometrics* 2003; **59**:701–709.

DOI: 10.1002/sim