Biometrics 59, 701–709
September 2003

Incorporating Data Received after a Sequential Trial Has Stopped into the Final Analysis: Implementation and Comparison of Methods

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

[1]Department of Statistics, University of Colombo, Sri Lanka
[2]Medical and Pharmaceutical Statistics Research Unit, University of Reading, U.K.
[*]*email:* j.r.whitehead@reading.ac.uk

Summary. In a sequential clinical trial, accrual of data on patients often continues after the stopping criterion for the study has been met. This is termed "overrunning." Overrunning occurs mainly when the primary response from each patient is measured after some extended observation period. The objective of this article is to compare two methods of allowing for overrunning. In particular, simulation studies are reported that assess the two procedures in terms of how well they maintain the intended type I error rate. The effect on power resulting from the incorporation of "overrunning data" using the two procedures is evaluated.

Key words: Clinical trials; Delayed data; Interim analysis; Overrunning; P-value function; Sequential methods.

1. Introduction

In many sequential clinical trials valid data on the primary efficacy outcome continue to be collected after the stopping criterion for the study has been reached. This phenomenon is termed "overrunning." The additional data are referred to as "overrunning data." Scientifically, data from as many as possible of the patients randomized into a clinical trial should be included in the final analysis (ICH Guideline E9, 1998, Section 5.2.1, `http://www.ifpma.org/pdf.fpma/e9.pdf`, International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1998). However, the overrunning data should be included only if they are "valid" in the sense of having been collected according to the protocol and such that the treatment and assessment of the patients concerned were not influenced by the fact that a stopping criterion had been met. This article discusses ways in which overrunning data can be incorporated into the final analysis of the trial.

A method for incorporating overrunning data into a final trial analysis was proposed by Whitehead (1992) and is implemented in the software PEST 4 (MPS Research Unit, 2000). It has been used in the analysis of numerous sequential trials (Whitehead, 1993; Moss et al., 1996; Derry et al., 1997). The method involves ordering the potential final datasets in the manner suggested by Fairbanks and Madsen (1982). Experience with the analysis of Moss et al. (1996) prompted Hall and Ding (2001) to propose a new approach based on a method of combining p-values which was then implemented in Moss et al. (2002). In this article, these two approaches will be described, evaluated, and compared.

It should be stressed that in both of the methods considered in this article, it is envisaged that once the stopping criterion has been reached, recruitment to the study will be stopped, and that subsequent reopening of recruitment will not be an option. This is a practical and widespread convention, although some individual clinical trials might operate differently. This convention means there is no point at the time of the overrunning analysis considering whether the original stopping criterion would still hold: there will be no further data.

The trial of Viagra in erectile dysfunction (Derry et al., 1997) provides a simple example of overrunning. Men with erectile dysfunction due to spinal cord injury, and with a regular female partner, were randomized between Viagra and placebo. The primary outcome was the patient's subjective assessment of erectile improvement after one month of treatment, expressed as success or failure. A triangular test was adopted and the sample path crossed the upper boundary (indicating superiority of Viagra) after results from 20 patients had become available. Recruitment to the trial was stopped, but at this stage, there were 6 men who had been randomized to treatment, but who had not yet provided their one-month assessment. As there was no reason to suppose that their assessments of treatment would be affected by the termination of the trial, it was felt to be both valid and appropriate to

include them in the final analysis. Further details of how this was done are given in Section 4.

A second example of overrunning is provided by the ASCLEPIOS (A Sandoz Clinical Evaluation Program of Israpidine on Stroke) study (Whitehead, 1993). This was a clinical comparison of an experimental calcium channel blocker with a placebo control in the immediate treatment of patients who had suffered an acute ischaemic stroke. The primary outcome variable was the Barthel index (Mahoney and Barthel, 1965), assessed 90 days after randomization. This measure of the patient's functional status is expressed as an ordinal outcome, with 21 possible values ranging from 0 (completely vegetative) to 100 (complete recovery) in steps of 5. An additional state (which can be thought of as −5) was added to the scale to represent death within 90 days.

Recruitment began in October 1989 and the first interim analysis was conducted in September 1990 on data from 140 patients. The sample path had crossed the lower boundary, indicating termination of the study due to the ineffectiveness of the study drug. An independent review panel confirmed this decision and recruitment to the study was stopped. Follow-up of all randomized patients to day 90 was continued. Thirteen months later, in October 1991, a final analysis was conducted on data from 229 patients, which included all patients recruited during the 90 days preceding the termination of the study, as well as those whose records were incomplete at the time of termination. As the treatment period lasted only a few days, most of these patients had already received all of their study medication when recruitment to the study was stopped. There was no way of altering their treatment, and the blinding of the study was not broken. It was felt that their responses would not be affected by the closure of the trial's recruitment, and so they were incorporated into the final analysis. In this example, the increase in sample size between the termination of recruitment and the final analysis was large, in part due to delays in data transfer. The lengthy delay before the final analysis was due to the process of validating all of the data.

2. Methods of Incorporating Overrunning Data

2.1 *Analysis with No Overrunning*

The methods of incorporating overrunning data discussed in this article are extensions of the method of analyzing data from a sequential trial without overrunning, which is described in Section 5.4 of Whitehead (1997), and incorporated into the software packages PEST (MPS Research Unit, 2000), EaSt 2000 (Cytel, 2000), and SeqTrial (MathSoft, 2000). That method will be outlined here, using notation from the first of these references.

Patients are randomized between an experimental treatment (E) and a control (C), and the parameter θ denotes a measure of the advantage of E over C. The trial can be expressed as a test of the null hypothesis H_0: $\theta = 0$. Interim analyses can be conducted in terms of the statistics Z and V, where Z denotes the efficient score representing the cumulative observed superiority of E over C, and V denotes Fisher's observed information, the information about θ gathered so far. A sequential trial comprises a number of interim analyses, at the ith of which the current score statistic Z_i is plotted against the current value of Fisher's information, V_i. Upper and lower stopping boundaries, denoted, respectively, by u_1, $u_2, \ldots$ and ℓ_1, $\ell_2, \ldots$ are defined in the Z-V plane. If $Z_i \geq u_i$ then the trial will stop, with the conclusion that E is superior to C. If $Z_i \leq \ell_i$ then the trial will stop, with the conclusion that E is inferior to or no different from C, depending on the design used and at what interim analysis this occurs. For designs with a maximum intended value of Fisher's information $V_{\max}$, stopping will also occur if $V_i \geq V_{\max}$.

Suppose that the trial terminates at the Tth inspection, where T is a random variable. Then, either Z_T lies outside the interval (ℓ_T, u_T) or $V_T \geq V_{\max}$. Denoting observed values of T and Z_T by t and z_t, respectively, the outcome (T, Z_T) is considered to be *more extreme than* (t, z_t) *if* $(T < t$ and $Z_T \geq u_T)$ *or* $(T = t$ and $Z_T \geq z_T)$. This ordering is due to Fairbanks and Madsen (1982). The p-value function $P(\theta)$ can be defined by $P(\theta) = P\{(T, Z_T)$ is more extreme than $(t, z_t); \theta\}$. It can be used to derive a one-sided p-value p^+, a median unbiased estimate θ_M, and a $100(1-\alpha)\%$ confidence interval (θ_L, θ_U) from $p^+ = P(0)$, $P(\theta_M) = 0.5$, $P(\theta_L) = \frac{1}{2}\alpha$, and $P(\theta_U) = 1 - \frac{1}{2}\alpha$, respectively. A one-sided p-value relating to the alternative H_1^-: $\theta < 0$ is given by $p^- = 1 - p^+$, and the two-sided p-value is $p = 2\min(p^+, p^-)$. The p-value function can be computed from

$$P(\theta) = P\{(T < t, Z_T \geq u_T) \text{ or } (T = t, Z_T \geq z_T); \theta\} \quad (2.1)$$

and the joint distribution of T and Z_T; it will be monotonically increasing.

Notice that if the trial stops at the first inspection, the analysis just described will be identical to the naive fixed sample size analysis conducted on the available data.

2.2 *The Deletion Method*

The "deletion method" is the name that will be given here to the approach introduced by Whitehead (1992), as its essential feature is that the interim analysis leading to the recruitment stoppage is deleted in the final analysis; otherwise, it is like any other application of the Fairbanks and Madsen ordering.

Suppose that the sample path reaches a stopping boundary at the tth interim analysis, but that data continue to be collected and there is another inspection. The deletion method analyses the trial as if the only interim analyses to have taken place were those for $I = 1, \ldots, t-1$ and $t+1$. That is, the tth interim analysis is effectively deleted from the record. Equation (2.1) becomes

$$P(\theta) = P\{(T < t, Z_T \geq u_T) \text{ or } (T = t+1, Z_T \geq z_T); \theta\}. \quad (2.2)$$

This is evaluated assuming that the design only allowed stopping with $V = V_1, \ldots, V_{t-1}$ or V_{t+1}. There is no need to assign values to ℓ_{t+1} or u_{t+1} for the calculation of $P(\theta)$. The deletion method reduces to the analysis described in Section 2.1 when $V_t = V_{t+1}$ and $Z_t = Z_{t+1}$.

2.3 *The Method of Combining p-values*

The method due to Hall and Ding (2001) will be referred to here as "the method of combining p-values." Suppose that $P_1(\theta)$ and $P_2(\theta)$ are two monotonically increasing p-value functions, relating to different datasets, but to the same parameter θ, and that w_1 and w_2 are constants satisfying $w_1^2 + w_2^2 = 1$. Hall and Ding show that $P(\theta)$ defined as

$$P(\theta) = 1 - \Phi[w_1 g\{P_1(\theta)\} + w_2 g\{P_2(\theta)\}] \quad (2.3)$$

is also a monotonically increasing p-value function, in the sense that it will give rise to valid p-values and confidence limits when used as described in Section 2.1, where Φ denotes the standard normal distribution function and $g(x) = \Phi^{-1}(1-x)$ for all $x \in (0, 1)$.

Set $V_O = V_{t+1} - V_t$ and $Z_O = Z_{t+1} - Z_t$. Define $P_1(\theta)$ to be the p-value function arising from the sequential portion of the trial as defined in equation (2.1), and $P_2(\theta)$ to be the p-value function arising from the overrunning data, treated as if from an independent fixed-sample trial. Making asymptotic normal assumptions, that is $P_2(\theta) = 1 - \Phi\{(Z_O - \theta V_O)/(V_O)\}$. These definitions, together with equation (2.3), allow an analysis of the trial that incorporates overrunning for any pair of weights w_1 and w_2. Hall and Ding use the weights

$$w_1 = \sqrt{\frac{V_T}{V_T + V_O}} \quad \text{and} \quad w_2 = \sqrt{\frac{V_O}{V_T + V_O}}, \qquad (2.4)$$

where V_T is the value of Fisher's information at the trial's termination. They point out that these are not constant so that, although the p-value function given by (2.3) will be monotonically increasing, it may not lead to valid p-values or confidence limits. They use some precise computations to demonstrate that any lack of validity is likely to be small. In this article, w_1 and w_2 given by equation (2.4) are referred to as "the random weights."

Hall and Ding also discuss the use of expected values of V in the definition of weights. Expected values of V_T and V_O could be used, but the properties of sample sizes (or in the case of survival data, of numbers of events) are likely to be more readily accessible in practice, and expectations in terms of V are likely to be approximately proportional to those in terms of the sample size n. Here, we will explore the weights

$$w_1' = \sqrt{\frac{E(n_{T;0})}{E(n_{T;0}) + E(n_{O;0})}} \quad \text{and}$$

$$w_2' = \sqrt{\frac{E(n_{O;0})}{E(n_{T;0}) + E(n_{O;0})}}, \qquad (2.5)$$

where n_T and n_O denote the sample sizes in the sequential and overrunning portions of the trial, respectively, and expectations are computed under the null hypothesis. In this article, w_1' and w_2' given by equation (2.5) are referred to as "the fixed weights," and as they are constants, the result of Hall and Ding (2001) we discussed following equation (2.3) above applies. Thus, this second choice will lead to valid p-values, although in practice, the relative weightings of the two portions of information might not be the most appropriate resulting in an inefficient analysis. When using the fixed weights, their values should be fixed prior to starting the trial, or at least prior to conducting the first interim analysis. In reality, the value used for $E(n_O; 0)$ is likely to be based on guesswork, probably anticipating some constant amount of overrunning. Inaccuracy in anticipating the value of $E(n_O; 0)$ will further erode the efficiency of the weighting, but it will not compromise the validity of the analysis.

3. Examples of Analyses Incorporating Overrunning

3.1 *The Trial of Viagra*

In the Viagra trial, the target treatment advantage was an increase in success rate from $p_C = 0.25$ on control to $p_E = 0.60$ on experimental, which corresponded to a log-odds ratio of $\theta = 1.50$. The power was set at 0.80 to detect $\theta = 1.50$ as significant at the two-sided 0.05 level. The trial was conducted as a triangular test, with stopping boundaries at $Z = (-2.834 + 1.586\,V)$ and $Z = (2.834 + 0.529\,V)$, adjusted for discrete looks using the "Christmas tree correction" (Whitehead, 1997). The first look was designed to take place after responses were available from 12 patients, and subsequent looks after every 4 new responses. At the design stage, it could be seen that $E(n_T; 0) = 38.1$; as the anticipated recruitment rate was one patient per week, it is reasonable to set $E(n_O; 0) = 4$ for use in setting the fixed weights.

Denote by n_C and n_E the number of responses and by S_C and S_E the number of successes, on C and E, respectively. Set $n = n_C + n_E$, $S = S_C + S_E$, and $F = n - S$. Then the test statistics are given by $Z = (n_C S_E - n_E S_C)/n$ and $V = n_E n_C SF/n^3$. The first three interim analyses took place after 12, 16, and 20 responses had been received. Values for (V, Z) were (0.750, 2.000), (0.984, 2.500), and (1.238, 3.500), respectively. The third point lay above the upper stopping boundary, so recruitment was stopped. However, 6 patients were still under treatment at this stage.

Once the data from these 6 patients were available, an overrunning analysis to incorporate them was conducted, resulting in $V = 1.529$ and $Z = 4.385$. This point remained above the upper stopping boundary. Table 1 presents the p-values (two-sided alternative), median unbiased estimates of θ and 95% confidence intervals, from various analyses based on these data.

The fixed weights given by (2.5) were $w_1' = 0.951$ and $w_2' = 0.308$, whereas the random weights given by (2.4) turned out to be $w_1 = 0.900$ and $w_2 = 0.437$. This indicates that use of the fixed weights in this study takes too much account of

Table 1

Alternative final analyses of the Viagra and ASCLEPIOS trials; the methods are 0: *ignoring overrunning,* 1: *deletion method,* 2: *combining p-values* (*random weights*) *and* 3: *combining p-values* (*fixed weights*)

	Viagra			ASCLEPIOS		
Method	p	θ_M	(θ_L,θ_U)	p	θ_M	(θ_L,θ_U)
0	0.00377	2.735	(0.906, 4.527)	0.225	−0.382	(−0.998, 0.235)
1	0.00313	2.718	(0.972, 4.362)	0.678	−0.099	(−0.569, 0.370)
2	0.00089	2.794	(1.164, 4.401)	0.678	−0.099	(−0.569, 0.370)
3	0.00111	2.777	(1.128, 4.401)	0.466	−0.180	(−0.663, 0.304)

the sequential part of the trial. The two combining p-values approaches have resulted in the smallest p-values and the largest point estimates.

3.2. *The ASCLEPIOS Trial*

In the ASCLEPIOS study, the primary outcome variable was the Barthel index (or death) at 90 days. The proportional-odds model (McCullagh, 1980) was assumed, and the common log-odds ratio θ on better outcomes adopted as the measure of treatment advantage. The power was set at 0.90 to detect a target improvement on this scale of 0.56 as significant at the two-sided 0.05 level. A log-odds ratio of $\theta = 0.56$ corresponded to an approximate reduction in the 90-day death rate from 15% on control to 9% on experimental, and to a commensurate improvement in the other Barthel categories. The trial was conducted as a triangular test, with stopping boundaries at $\mathrm{Z} = (-8.809 + 0.510\,\mathrm{V})$ and $\mathrm{Z} = (8.809 + 0.170\,\mathrm{V})$.

The anticipated recruitment rate was 15 patients per month. The first interim analysis was planned to take place when 140 patients had completed their 90-day evaluations which, allowing for the three-month delay in response, was anticipated at 12 months after the start of the study. Subsequent interim analyses were planned after every 90 new responses, anticipated to be at six monthly intervals. For this design $\mathrm{E}(n_\mathrm{T};\ 0) = 236$. As the anticipated recruitment rate was 15 patients per month, allowing for the three-month delay in response plus one month for data transfer, it was reasonable to set $\mathrm{E}(n_\mathrm{O};\ 0) = 60$ for use in setting the fixed weights.

The test statistics Z and V were calculated with adjustment for the neurological score at baseline, time from stroke to medication, and geographic region, as described by Whitehead (1993, 1997) and implemented in PEST 4.

Recruitment began in October 1989, and the first interim analysis was performed in September 1990, when 90-day Barthel indices were available for 140 patients. Values for V and Z were (10.104, −3.855); the lower stopping boundary had been crossed and recruitment to the trial was terminated. Follow-up of patients continued until their 90-day Barthel index was recorded. An overrunning analysis to incorporate these data from 89 additional patients was conducted in October 1991. The resulting values were V = 17.410 and Z = −1.728, and this point remained below the stopping boundary. The results from various analyses are included in Table 1. The fixed weights given by (2.5) were $w_1' = 0.893$ and $w_2' = 0.450$, whereas the random weights given by (2.4) turned out to be $w_1 = 0.762$ and $w_2 = 0.648$. In this case, the deletion method and the random-weights approach give identical analyses. This is because, as pointed out at the end of Section 2.1, the analysis based on the Fairbanks and Madsen (1982) ordering coincides with the fixed-sample analysis when stopping occurs at the first interim analysis. The deletion method deletes the first interim analysis, leaving a trial with only one analysis. The random weights method combines a p-value from a one-interim analysis test with a fixed-sample overrunning analysis, leading back to that same fixed-sample analysis. The fixed-weights approach underweights the slight revival of experimental fortunes during the overrunning phase, and thus gives results closer to the analysis that ignores overrunning.

4. A Simulation Study

A simulation study was conducted in the setting of the ASCLEPIOS study. Day-90 Barthel indices were simulated for each patient in the study on a six-point scale representing the outcomes death, score 0, scores 5–35, 40–65, 70–95, or 100. For control patients, the probabilities of these outcomes were taken to be 0.169, 0.015, 0.242, 0.318, 0.181, and 0.075, respectively, as used at the design stage (Whitehead, 1993). (To overcome rounding errors, the first probability has been increased from 0.166 to 0.169 for use here.) The trial was simulated both under H_0 and under a proportional-odds alternative (H_1) with log-odds ratio $\theta = 0.56$; for H_1 the corresponding six outcome probabilities for experimental patients were 0.104, 0.010, 0.184, 0.326, 0.252, and 0.124, respectively.

For all simulation runs, the two-sided significance level was set to be $\alpha = 0.05$ and the power when $\theta = 0.56$ to be $1 - \beta = 0.90$. The same triangular design as used in the real trial was examined, together with a restricted procedure with slope zero satisfying the same specification (Whitehead, 1997). This procedure coincides with the O'Brien and Fleming (1979) design. However, in setting the stopping limits, the PEST 4 implementation uses the Christmas tree correction to allow for the intervals between interim analyses, rather than the more accurate approach of recursive numerical integration. In final analyses, the recursive numerical integration method can be adopted.

Two analysis inspection intervals (every 30 or every 90 new patient responses) were explored. The latter choice matched the planned analysis inspection interval for the actual trial (following the first interim analysis) and the former was included for the simulations of the triangular test only. It was imagined that the *anticipated* overrunning would comprise 60 new patient responses. Most of the simulations investigate the case in which 60 overrunning patients were actually observed, but some of those for the triangular test included a scenario in which the overrunning was actually twice that anticipated, giving 120 new-patient responses. For comparison, situations in which no overrunning was anticipated and none was observed were also considered. Ten thousand replicate simulations were conducted for each setting.

For the triangular test using the method of combining p-values with fixed weights, inspection intervals of 30 and 90 correspond to $\mathrm{E}(n_\mathrm{T};\ 0) = 221$ and 236, respectively. The anticipated amount of overrunning was 60 patients, and so $\mathrm{E}(n_\mathrm{O};\ 0) = 60$ was used in (2.5) to calculate w_1' and w_2'. For the O'Brien and Fleming design, $\mathrm{E}(n_\mathrm{T};\ 0) = 444$.

Tables 2–5 present the principal results for the triangular test. For each situation, the following items are recorded. First, the counts of trials crossing the upper boundary are given. (These should be equal to 250 under H_0 and 9000 under H_1.) For ease of comparison, the simulations without overrunning reported in Table 2 were constructed to match the later runs performed with an overrun of 60 patients. This was accomplished by using the same random seed and generating the overrunning data, but not using them. Next are given the number of times the one-sided p-value for evidence favoring the experimental treatment (p^+) is less than 0.0125 and is less than 0.025, and the number of times the one-sided p-value for evidence favoring the control treatment (p^-) is less than 0.025.

Table 2
Simulation results for the triangular test with no overrunning

Hypothesis	H_0		H_1	
Inspection interval	30	90	30	90
Crossed upper	249	261	8981	8992
$p^+ \le 0.0125$	116	125	6430	6753
$p^+ \le 0.025$	249	260	8981	8992
$p^- \le 0.025$	232	217	1	1
$\theta_L > \theta$	249	260	205	244
$\theta_M > \theta$	4972	5042	4957	4981
$\theta_U > \theta$	9768	9783	9729	9766
P97.5(θ_L)	−0.0002	0.0018	0.5356	0.5563
MED(θ_M)	−0.0016	0.0023	0.5573	0.5588
P02.5(θ_U)	0.0158	0.0209	0.5538	0.5671

The number of times the 95% confidence limits θ_L and θ_U and the median unbiased estimate θ_M exceed the value of θ used in the simulation are then given (cf. 250, 5000, and 9750 for θ_L, θ_M, and θ_U, respectively). Finally, the 97.5th percentile of θ_L, the median of θ_M, and the 2.5th percentile of θ_U (respectively, denoted by P97.5(θ_L), MED(θ_M), and P02.5(θ_U)), are given. These should be equal to the value of θ used in the simulation. Based on 95% probability intervals for proportions, counts with a target of 5000 can be expected to lie about 100 either side; those with a target of 9000 should be within 60 either side; those with a target of 250 or 9750 should be within 31 either side; and those with a target of 125 should be within 22 either side.

Table 2 gives satisfactory results for the Fairbanks and Madsen (1982) ordering, in the absence of overrunning. There is not an exact correspondence between crossing the upper boundary and finding that $p^+ \le 0.025$, because the latter takes into account the overshoot of the boundary at the first interim analysis in which the boundary is crossed. Discrepancies are particularly likely when stopping is very late and p^+ is very close to 0.025. The upper limit of the confidence interval tends to be on the large (conservative) side when the inspection interval is 90 patients. Table 3 shows that, for the most part, the method of combining p-values based on fixed weights achieves the target values satisfactorily. For an inspection interval of 90 and under H_1, the power to detect $p^+ \le 0.0125$ is 78% for an overrun of 60 and 85% for an overrun of 120, illustrating the ability of this method to use the overrunning data to strengthen positive conclusions. However, the principal power

Table 3
Simulation results for the triangular test, with method of combining p-values with fixed weights

Hypothesis	H_0				H_1			
Inspection interval	30		90		30		90	
Overrun	60	120	60	120	60	120	60	120
Crossed upper	249	261	261	250	8981	8984	8992	8998
$p^+ \le 0.0125$	113	141	125	120	7626	8484	7795	8507
$p^+ \le 0.025$	228	259	256	251	8664	9111	8751	9158
$p^- \le 0.025$	240	243	269	232	1	0	0	0
$\theta_L > \theta$	228	259	256	251	214	248	243	224
$\theta_M > \theta$	4979	5027	5074	4983	4949	5021	5008	4973
$\theta_U > \theta$	9760	9757	9731	9768	9734	9753	9762	9758
P97.5(θ_L)	−0.0056	0.0033	0.0017	0.0006	0.5476	0.5595	0.5532	0.5480
MED(θ_M)	−0.0011	0.0014	0.0042	−0.0008	0.5574	0.5610	0.5604	0.5574
P02.5(θ_U)	0.0074	0.0025	−0.0081	0.0088	0.5526	0.5616	0.5635	0.5615

Table 4
Simulation results for the triangular test, with method of combining p-values with random weights

Hypothesis	H_0				H_1			
Inspection interval	30		90		30		90	
Overrun	60	120	60	120	60	120	60	120
Crossed upper	249	261	261	250	8981	8984	8992	8998
$p^+ \le 0.0125$	102	114	106	99	7582	8277	7771	6884
$p^+ \le 0.025$	212	242	241	236	8742	9002	8828	9130
$p^- \le 0.025$	229	222	216	217	1	0	0	0
$\theta_L > \theta$	212	242	241	236	196	222	198	79
$\theta_M > \theta$	5150	5287	5239	5205	4777	4763	4840	4219
$\theta_U > \theta$	9771	9778	9784	9783	9770	9812	9791	9860
P97.5(θ_L)	−0.0089	−0.0044	−0.0029	−0.0044	0.5288	0.5449	0.5358	0.4665
MED(θ_M)	0.0082	0.0099	0.0132	0.0099	0.5494	0.5490	0.5526	0.5111
P02.5(θ_U)	0.0099	0.0143	0.0163	0.0143	0.5648	0.5761	0.5764	0.6143

Table 5
Simulation results for the triangular test, with deletion method

Hypothesis	H_0				H_1			
Inspection interval	30		90		30		90	
Overrun	60	120	60	120	60	120	60	120
Crossed upper	249	261	261	250	8981	8984	8992	8998
$p^+ \leq 0.0125$	66	60	74	52	6144	6201	6798	6884
$p^+ \leq 0.025$	214	204	205	180	8935	9088	8968	9130
$p^- \leq 0.025$	145	114	170	138	1	0	0	0
$\theta_L > \theta$	214	204	205	180	100	84	124	79
$\theta_M > \theta$	5625	6135	5784	6032	4325	4013	4605	4219
$\theta_U > \theta$	9855	9886	9830	9862	9835	9890	9841	9860
P97.5(θ_L)	−0.0032	−0.0044	−0.0078	−0.0090	0.4707	0.4630	0.4800	0.4665
MED(θ_M)	0.0431	0.0592	0.0328	0.0384	0.5238	0.5107	0.5345	0.5111
P02.5(θ_U)	0.1151	0.1189	0.0616	0.1051	0.5970	0.6201	0.6079	0.6143

requirement, that for detecting $p^+ \leq 0.025$, is estimated as only 87.5% for an overrun of 60. The corresponding count is 249 below the target of 9000, which is further away than is consistent with chance. The situation is satisfactory for an overrun of 120, and indeed there is a gain of power.

The method of combining p-values with random weights is acknowledged to lead to invalid analyses, and this is reflected in the results in Table 4. The analyses are too conservative, with the type I error rates appearing to be below the set 0.025 in each direction, especially for negative conclusions. The confidence intervals tend to be too wide, and under H_1, the median unbiased estimate is too small. For an inspection interval of 90 and under H_1, the power to detect $p^+ \leq 0.0125$ is 78% for an overrun of 60, but only 69% for an overrun of 120. The former matches the fixed-weights case, as the overrun is 60 as anticipated. The latter is poorer because the overrun is given proportional weighting in the random-weights analysis, while being constrained to what is an underrepresentation in the fixed-weights case. Once more, the power to detect $p^+ \leq 0.025$ is reduced, this time to 88.3% for an overrun of 60. Again, there is a gain of power for an overrun of 120.

Table 5 shows that the deletion method departs even further from target values than the method of combining p-values with random weights. The conservatism is greater and the power to detect $p^+ \leq 0.0125$ is less, but the loss of power to detect $p^+ \leq 0.025$ is only very small.

When the amount of overrun is 60 patients, all three methods show some loss of power to detect $p^+ \leq 0.025$ as a result of incorporating the overrunning data. This is because when the upper boundary is crossed, an analysis without the overrunning data would almost certainly give $p^+ \leq 0.025$. Adding the overrunning data can improve estimates and lead to a lower p-value, but in terms of finding that $p^+ \leq 0.025$, it can only make matters worse. If the lower boundary is crossed, then overrunning data can redeem the situation, but this becomes likely only when there is a large overrun.

Table 6 further investigates loss of significance. The upper portion presents further data from the runs reported in Tables 2, 4, and 5, in the case of an inspection interval of 90 and an overrun of 60. All simulated trials in which the final conclusion was inconsistent with the boundary originally crossed were classified by the interim analysis at which recruitment was terminated (that is, the one before the final analysis incorporating overrunning data). The method of combining p-values with fixed weights leads to the greatest number of changes in conclusion, although reversals of conclusions drawn after trials stopped very early (that is, at the first or second interim analysis) are unusual. The deletion method has by far the least number of reversals, with the random-weights method lying in between the other two methods, but closer to the fixed-weights case. To create the lower portion of Table 6, 10,000 replicate simulations were run under H_1 using an inspection interval of 90, together with nine different amounts of overrunning data. The three analysis methods were then applied, and in each case, the fixed weights were based on the anticipated overrun of 60 observations. The number of cases in which significance was lost (upper boundary crossed but final $p^+ > 0.025$) and the number in which significance was gained (lower boundary crossed but final $p^+ \leq 0.025$) were counted. Loss of significance is most common when the method of combining p-values with fixed weights is used, and least common for the deletion method.

The results for the O'Brien and Fleming test are presented in Table 7. These present a slightly different picture. First, even in the no-overrunning case, there is a noticeable discrepancy between crossing the upper boundary and finding that $p^+ \leq 0.025$, the latter being less likely under either hypothesis. It is only possible for crossing of the upper boundary to be followed by an analysis yielding $p^+ \leq 0.025$ if the crossing takes place at the fifth and final possible interim analysis. As mentioned earlier, this is because the implementation in PEST 4 calculates stopping limits using the Christmas tree correction, as opposed to computing p-values and confidence limits using recursive numerical integration. As shown by Stallard and Facey (1996), the Christmas tree correction is extremely accurate for the triangular test, but less so for other designs such as restricted procedures.

For the no-overrunning case, the analysis after the O'Brien and Fleming design adheres closely to theoretical predictions. When overrunning occurs, the three methods are in closer agreement with one another than in the case of the triangular test. All three are conservative, and all three lose power to

Table 6

Simulation results for the triangular test under H_1 with an inspection interval of 90*—changes of conclusion by stopping inspection and by amount of overrunning*

	Combining p-values with fixed weights		Combining p-values with random weights		Deletion	
	Upper and $p^+ \geq 0.025$	Lower and $p^+ \leq 0.025$	Upper and $p^+ \geq 0.025$	Lower and $p^+ \leq 0.025$	Upper and $p^+ \geq 0.025$	Lower and $p^+ \leq 0.025$
Stopping inspection	**Number of runs (out of 10,000) stopping at the inspection indicated and changing conclusion at the final analysis (overrun = 60)**					
1	0	0	1	0	1	0
2	14	0	32	0	25	0
3	111	5	94	4	46	2
4	174	48	128	38	35	19
5	150	108	115	98	22	41
6	60	96	47	102	16	61
7	5	16	5	16	5	3
Total	514	273	422	258	150	126
Overrunning	**Total number of runs (out of 10,000) changing conclusion at the final analysis when the amount of overrunning is as indicated**					
0	4	0	4	0	4	0
15	1413	139	363	87	66	14
30	954	200	426	151	108	48
45	757	234	471	201	155	95
60	514	273	422	258	150	126
75	448	292	423	280	155	156
90	343	323	422	324	173	214
105	289	358	408	377	173	242
120	232	368	345	398	154	286

detect significance with either $p^+ \leq 0.0125$ or $p^+ \leq 0.025$ relative to the no-overrunning case. The fixed-weights method does not adhere to theory as closely as when used for the triangular test, and this is partly because the weights are no longer completely fixed. This occurs because recruitment is closed after 450 patients. If the trial continues to the fifth and final possible analysis, then there will be no overrunning data. The algorithm used in the simulations then returns the standard, no-overrunning, analysis. So in these circumstances, the weights used are actually $w_1' = 1$ and $w_2' = 0$. Continuation to the very end is common with this design, and so the fixed- and random-weights approaches become similar to one another. The deletion method still suffers the least loss of power.

In these simulations, no trial that stopped on the lower boundary ever reached a positive significant conclusion, due to the addition of overrunning data. Clearly, this is more difficult to do with an O'Brien and Fleming design than with a triangular test. Stopping early on the upper boundary, followed by a nonsignificant result when overrunning data were incorporated, was also rare, occuring at the fourth interim analysis only, 3 times with the fixed-weights method, 17 times

Table 7

Simulation results for the O'Brien and Fleming test, inspection interval of 90*, no overrunning or overrunning of* 60

Method	No overrunning		Comb. p, fixed		Comb. p, random		Deletion	
Hypothesis	H_0	H_1	H_0	H_1	H_0	H_1	H_0	H_1
Crossed upper	276	9055	276	9055	276	9055	276	9055
$p^+ \leq 0.0125$	115	7554	84	7440	76	7335	80	7431
$p^+ \leq 0.025$	252	8988	234	8935	225	8900	232	8939
$p^- \leq 0.025$	250	0	221	0	209	0	217	0
$\theta_L > \theta$	252	226	234	226	225	216	232	207
$\theta_M > \theta$	5017	5042	5017	5062	5017	4983	5017	4458
$\theta_U > \theta$	9750	9762	9779	9762	9791	9762	9783	9762
P97.5(θ_L)	0.0002	0.5486	−0.0046	0.5500	−0.0075	0.5469	−0.0050	0.5409
MED(θ_M)	0.0010	0.5620	0.0010	0.5636	0.0010	0.5592	0.0010	0.5422
P02.5(θ_U)	0.0000	0.5630	0.0065	0.5630	0.0085	0.5630	0.0073	0.5630

with the random-weights method, and 13 times with the deletion method.

5. Conclusions

In order for sequential designs to be applied to clinical trials with the commonly occurring feature of delayed patient responses, it is necessary to have a strategy for dealing with overrunning data. Simply ignoring this additional information would be convenient, but it is not scientifically appropriate and it runs counter to regulatory advice. Surprisingly, inclusion of the overrunning data may not increase the power of the trial, and indeed it appears to be possible for an overrunning analysis to be less powerful than an analysis ignoring the extra data. This is because early stopping on the upper boundary implies that a significant result has already been found, and extra data can then only lose this positive result. For the triangular test, this consideration can be outweighed by apparently nonsignificant trials becoming significant after the addition of the overrunning data, but such behavior is far less likely for the O'Brien and Fleming design. In the case of the triangular test, the power to achieve a more persuasive level of significance can be enhanced through the incorporation of overrunning data.

Choosing between the methods for analyzing the final dataset is based on two criteria. The first is the reliability of the analysis produced, with accuracy being most desirable, and conservatism being acceptable. The second is the stability of the findings, with loss of significance between the time of stopping recruitment and the final analysis being a major disadvantage. This is because the imposition of a sequential design would encourage investigators to stop in view of apparently positive conclusions, and yet they will be left with a nonsignificant trial. The conclusion that a larger fixed-sample trial design would have been preferable will then be hard to avoid. When recruitment closes late in the trial, with the sample size close to or exceeding the equivalent fixed sample size, this "regret" will be small, but the consequences of stopping a trial very early, only to end without a positive result, are far more serious.

When the fixed-weights method of combining p-values can be conducted, it provides very accurate results. The deletion method leads to the least accurate analyses, and for the triangular test, the power to obtain lower levels of significance than set is least enhanced. However, the analyses are conservative, and this method is the least likely to result in a change from a significant to a nonsignificant result. It would appear that the deletion method earns its stability by down-weighting the overrunning data in the final analysis. The method of combining p-values with random weights gives results lying between the properties of the other two approaches.

A further method has been suggested in the literature by Hall and Liu (2002), based on the maximum likelihood ordering suggested by Emerson and Fleming (1990). The major drawback with this method is that the maximum likelihood ordering is not "truncation adaptable," in the sense introduced by Liu and Hall (1999). This means that the analysis requires knowledge of the values of V_{T+1}, $V_{T+2}, \ldots$, at which interim inspections would have taken place had the trial not been stopped at the Tth interim analysis. Clearly, in applications, values have to be imputed for these quantities, but the principle of the method is not completely satisfactory. Among the methods explored in this article, and for the patterns of inspection and overrunning explored, the authors would recommend using the deletion method. This was not what we were expecting to find, as the extent of the loss of power and changes of conclusion inherent in the other methods had not been anticipated. These undesirable features seem to us to outweigh the improved accuracy when considering a method for practical use in clinical trials. In cases where interim analyses are more frequent than studied here, or where the amount of overrunning is likely to depend on the interim at which stopping occurs, the relative merits of the methods might be different.

The considerations highlighted here should also be taken into account when a design method is chosen. In particular, the option of delaying the first interim so that V_1 is considerably larger than the subsequent increments $(V_i - V_{i-1})$ is attractive, as this will avoid the situation of very early termination for treatment advantage that could be followed by an unfortunate reversal of fortune.

Acknowledgements

During this research, Dr Sooriyarachchi was in receipt of a Wellcome Trust Short-Term Travel Grant. The authors are grateful to Novartis Pharma for permission to include unpublished details of the ASCLEPIOS study, and to Jack Hall for helpful discussions.

Résumé

Dans un essai clinique séquentiel, l'accumulation de données sur les patients se poursuit souvent après que le critère d'arrêt pour l'étude ait été obtenu. Ceci est connu sous le terme de ≪dépassement≫ (≪overrunning≫). Le dépassement intervient principalement lorsque le critère principal de chaque patient est mesuré à la fin d'une période d'observation. L'objectif de cet article est de comparer deux méthodes de prise en compte de ce ≪dépassement≫. En particulier, on présente des études de simulation qui évaluent les deux procédures en fonction du contrôle du risque d'erreur de type I prédéfini. On évalue également l'effet sur la puissance résultant de l'incorporation de données de ≪dépassement≫ par chacune des deux procédures.

References

Cytel Software Corporation. (2000). *EaSt 2000: A Software Package for the Design and Interim Monitoring of Group Sequential Clinical Trials.* Cambridge, Massachusetts: Cytel.

Derry, F. A., Dinsmore, W. W., Fraser, M., Gardner, B. P., Glass, C. A., Maytom, M. C., and Smith, M. D. (1997). Efficacy and safety of oral sildenafil (Viagra) in men with eretile dysfunction caused by spinal cord injury. *Neurology* **51,** 1629–1633.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77,** 875–892.

Fairbanks, K. and Madsen, R. (1982). P values for tests using a repeated significance test design. *Biometrika* **69,** 69–74.

Hall, W. J. and Ding, K. (2001). *Sequential tests and estimates after overrunning based on p-value combination.* Technical Report 01/06, Department of Biostatistics, University of Rochester, Rochester, New York.

Hall, W. J. and Liu, A. (2002). Sequential tests and estimators after overrunning based on maximum-likelihood ordering. *Biometrika* **89,** 699–707.

Liu, A. and Hall, W. J. (1999). Unbiased estimation following a group sequential test. *Biometrika* **86,** 71–78.

Mahoney, F. I. and Barthel, D. W. (1965). Functional evaluation: The Barthel index. *Maryland State Medical Journal* **14,** 61–65.

MathSoft. (2000). *S-Plus.* Seattle: MathSoft.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42,** 109–142.

Moss, A. J., Hall, W. J., Cannom, D. S., et al., for the Multicenter Automatic Defribrillator Implantation Trial Investigators (1996). Improved survival with implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. *New England Journal of Medicine* **335,** 1933–1940.

Moss, A. J., Zareba, W., Hall, W. J., Klein, H., Wilber, D., Cannom, D. S., Daubert, J. P., Higgins, S. L., Brown, M. W., and Andrews, M. L., for the Multicenter Automatic Defribrillator Implantation Trial II Investigators (2002). Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New England Journal of Medicine* **346,** 877–883.

MPS Research Unit. (2000). *PEST 4: Operating Manual.* Reading, U.K.: University of Reading.

O'Brien, P.C. and Fleming. T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

Stallard, N. and Facey, K. M. (1996). Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics* **6,** 361–373.

Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials* **13,** 106–121.

Whitehead, J. (1993). Application of sequential methods to a phase III clinical trial in stroke. *Drug Information Journal* **27,** 733–740.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials,* revised 2nd edition. Chichester, U.K.: Wiley.

Received May 2002. *Revised January* 2003.
Accepted January 2003.