STATISTICS IN MEDICINE
*Statist. Med.* **17**, 2835–2847 (1998)

# SAMPLE SIZE REVIEW IN A HEAD INJURY TRIAL WITH ORDERED CATEGORICAL RESPONSES

K. BOLLAND[1]*, M. R. SOORIYARACHCHI[2] AND J. WHITEHEAD[1]

[1] *Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.*

[2] *Department of Statistics and Computer Science, The University of Colombo, P.O. Box 1490, Colombo, Sri Lanka*

## SUMMARY

Between 1993 and 1996, a total of 452 patients were entered into a randomized trial evaluating eliprodil (a non-competitive NMDA receptor antagonist) in patients suffering from severe head injury. The primary efficacy analysis concerned the Glasgow Outcome Score (GOS), six months after randomization. This outcome was classified into three ordered categories: good recovery; moderate disability, and the worst category made up by combining severe disability, vegetative state and dead. A sample size calculation was performed prior to the commencement of the study, using a formula which depends on the anticipated proportions of patients in the three different outcome categories, the proportional odds assumption and on the relationship between outcome and prognostic factors such as Glasgow Coma Score at entry. Owing to uncertainty about the influence of prognostic factors, and about the proportion of patients in the three GOS categories, a blinded sample size review was planned. This review was performed on the basis of the first 93 patients to respond, and this led to an increase in the sample size from 400 to 450. In this paper the pre-trial simulations showing that the type I error rate would not be influenced and the power would be preserved will be presented, and the implementation of the procedure will be described. © 1998 John Wiley & Sons, Ltd.

## 1. INTRODUCTION

Sample size calculations for clinical trials are usually conducted to determine the total number of patients needed. These calculations are made in an effort to satisfy a specified power requirement, and their validity is dependent on pre-trial modelling assumptions. This paper concerns the sample size calculation for a recently completed trial in head injury, and a mid-trial review driven by the uncertainty of these pre-trial modelling assumptions. The mid-trial review led to an increase in the number of patients recruited. Prior to conducting the review a simulation study was performed which confirmed that the review procedure would indeed meet its objective of preserving power without inflation of the type I error rate, and those simulation results are presented here.

The trial which forms the subject of this paper was a clinical comparison of the drug eliprodil with placebo. The primary response was an ordinal score measuring the functional status of the

* Correspondence to: K. Bolland, Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.

CCC 0277–6715/98/242835–13$17.50

patient after six months of treatment. Throughout the paper, attention is focussed on trials in which one experimental treatment (E) is compared with one control regimen (C) in terms of an ordered categorical response. The actual outcome of the trial of eliprodil will not be discussed here.

A standard form of power requirement specifies that a treatment difference should be found with probability $(1-\beta)$ if a true treatment difference equal to $\theta_R$ is present. By 'finding a treatment difference', obtaining a significant difference at the $100\alpha$ per cent level against the two-sided alternative is meant. The probability $(1-\beta)$ is referred to as the power of the test. The value $\theta_R$ represents a clinically relevant advantage of E over C and is a particular value of some measure $\theta$ of treatment difference. Sample size formulae and tables for a variety of types of endpoint are given by Machin *et al.*,[1] and the binary case is dealt with in the books by Pocock[2] and Fleiss.[3] Sample size calculations for survival data are given by many authors, including Freedman[4] and Collett.[5]

In this paper the sample size formulae for ordered categorical data given by Whitehead[6] will be used: alternative approaches include exact formulae[7] and a method based on simulation.[8] Limitations of the formulae of Whitehead for small samples and when model assumptions are violated have been pointed out by Kolassa[9] and Hilton,[10] respectively.

Most sample size formulae involve nuisance parameters, the values of which are unknown when the trial is designed. For example, with binary data, values for the success rate are needed for the formula: either the overall success rate, or individual values for E and C depending on which version is used. For survival data, assuming proportional hazards, the number of events required to achieve given power can be found without specification of nuisance parameters, but knowledge of the survival pattern is needed to translate these into sample sizes. With ordinal data, under proportional odds, it is the overall probability of each category that is required. Uncertainty about the values of nuisance parameters is the weakest link in the arguments which justify sample size, and is a common cause of underpowered and negative trials. It is this link which is examined for ordinal data in this paper.

The idea of a mid-trial sample size review goes back to work of Wittes and Brittain[11] and Gould.[12,13] These authors use the terms 'internal pilot study' and 'interim analysis', respectively. Here, the term 'sample size review' is preferred as it indicates the limited purpose of looking at the data, and distinguishes it from interim analyses at which treatments are compared with a view to stopping the trial. In a sample size review, data are examined in a blinded fashion, without treatment codes being attached. No indication of treatment difference is given, and so effects on the type I error rate are minimal. Simulation results have confirmed the latter statement with binary data.[12]

Previous papers have confined attention to the effects of re-estimation of nuisance parameters. In the case of binary and ordered categorical data, sample size formulae can be stratified.[6] When $\theta$ is interpreted as the treatment difference within strata, stratification leads to an increase in sample size. This feature is discussed for binary data by Robinson and Jewell[14] and for survival data by Ford *et al.*[15] The former authors show that the variance of the pooled estimate of the log-odds ratio is always less than or equal to that of the adjusted (stratified) estimate. The point estimate of the pooled treatment effect will tend to be smaller than the estimate from the adjusted model. They demonstrate, however, that it is more efficient to adjust for prognostic factors when testing for a treatment effect, despite the associated loss of precision in estimation. In this paper the sample size review is used to assess the influence of prognostic factors on outcome, and if necessary to introduce stratification into the sample size calculation.

 

The design of the trial of eliprodil in head injury, and the clinical problems which instigated the sample size review procedure, are described in Section 2. The plans for the review are presented in Section 3 together with the results of the simulation study. The implementation of the plan is described in Section 4, and Section 5 is a discussion of the wider suitability of these methods.

## 2. THE TRIAL OF ELIPRODIL IN SEVERE HEAD INJURY

### 2.1. Trial description

The primary objective of the study was to evaluate the efficacy and safety of eliprodil on the improvement of functional status six months after severe head injury. It was an international, multi-centre, phase III trial with a double-blind design. Patients were randomized to either eliprodil (a non-competitive NMDA receptor antagonist[16]) or placebo, in addition to the usual treatment in each centre. Eligible patients were those with closed head injury or skull base fracture with dural tear, with Glasgow Coma Scale (GCS) after non-surgical resuscitation between 4 and 8 inclusive, without eye opening. The maximum time limit between primary injury and administration of study drug was to be less than 12 hours.

Observation of each patient was for a total duration of 6 months, involving two phases. The treatment period lasted for 20 days, during which the study treatments were administered intravenously for 7 days, followed by oral or naso-gastric tube administration for 13 days. The follow-up period was for a further 160 days.

The primary efficacy criterion was the Glasgow Outcome Scale (GOS) at 6 months. The GOS comprises five ordered response categories assessing patients functional status: good recovery; moderate disability; severe disability; vegetative state, and dead. This measure was also assessed at days 21, 50 and 90. Important prognostic variables were thought to be the GCS at day 0, delay before administration of treatment, and age. The average recruitment rate anticipated was 30 patients per month.

### 2.2. Sample size calculation

For the primary efficacy analysis of the GOS at 6 months, data were to be modelled as ordered categorical data using the proportional odds regression model described by McCullagh,[17] and the treatment effect quantified using the log-odds ratio $\theta$ defined by equation (1) below. To avoid the possibility of claiming superiority of eliprodil in part because of its moving patients from 'dead' to 'vegetative', or 'vegetative' to 'severe disability', these categories were pooled. Hence, the primary efficacy criterion reduced the five category GOS to just three categories: good recovery (GR); moderate disability (MD), and severe disability/vegetative state/dead (SD/V/D). For data of this type, with no allowance for prognostic variables, use of the proportional odds regression model is equivalent to the Mann–Whitney test. The advantage of the modelling approach is that it enables adjustment for prognostic variables.

Data from previous studies in head injury suggested that the distribution of 6 month GOS in placebo patients would be 17 per cent, 30 per cent, and 53 per cent, respectively for the three categories GR, MD, and SD/V/D. For the conclusion of the trial to be considered clinically relevant it was necessary for the log-odds ratio to be such that the proportion of patients having one of the two better outcomes be increased from 47 per cent on placebo to 62 per cent on eliprodil: an improvement of 15 percentage points. The sample size calculation, and the eventual

 

proportional odds regression analysis, both proceed under the assumption of *proportional odds*, which can be explained as follows. Let $Q_{1E}$ and $Q_{2E}$ denote the probabilities of GR and (GR or MD), respectively, for patients receiving eliprodil (E). Define $Q_{1C}$ and $Q_{2C}$ similarly for placebo (C). Then the log-odds ratio $\theta$ is defined by

$$\theta = \log\left(\frac{Q_{jE}(1-Q_{jC})}{Q_{jC}(1-Q_{jE})}\right), \quad \text{for } j = 1, 2 \tag{1}$$

assumed the same for $j = 1$ and 2. The latter is the proportional odds assumption. A positive value of $\theta$ indicates superiority of eliprodil and a negative value indicates inferiority. The clinically relevant value of $\theta$ is denoted by $\theta_R$.

Under the assumption of proportional odds, the clinically relevant improvement of 15 percentage points mentioned above would arise from $\theta = 0{\cdot}610$, and the corresponding outcome distribution for eliprodil is 27·4 per cent, 34·6 per cent and 38·0 per cent for the categories GR, MD and SD/V/D. If this degree of treatment advantage in the eliprodil group, (which we will term the *reference improvement*) were present, then a power of 90 per cent was set to attain a result which was significant at the 5 per cent level (two-sided alternative). Let $\bar{p}_j$ denote the anticipated proportion of patients from both treatment groups falling into each category $j = 1, 2, 3$. Using the sample size formula of Whitehead,[6] the total number of patients required is $n$, where

$$n = \frac{12(u_{\alpha/2} + u_{\beta})^2}{\theta_R^2\left(1 - \sum_{i=1}^{3} \bar{p}_j^3\right)} \tag{2}$$

and $u_\gamma$ denotes the upper $100\gamma$ percentage point of the standard Normal distribution. Here $\bar{p}_1 = (0{\cdot}17 + 0{\cdot}274)/2 = 0{\cdot}222$, $\bar{p}_2 = 0{\cdot}323$ and $\bar{p}_3 = 0{\cdot}455$ giving a sample size of $n = 394$, which was rounded to 400 in the trial protocol.

The validity of the sample size calculation above depends on three assumptions. First, that the proportional odds model is at least approximately valid. Second, that the nuisance parameters involved in equation (2), namely the proportions of patients in each of the GOS categories on placebo, are at least approximately 0·17, 0·30 and 0·53. Third, that none of the prognostic factors recorded at baseline have a major effect on the 6 month GOS. If any of these assumptions do not hold then the sample size calculated will not be valid, and the trial may not achieve its target power.

Guesses for these nuisance parameters were not readily available. After a search of relevant literature of previous studies in head injury, appropriate values were assigned to the outcome probabilities $p_{jC}$ for the control group. Obviously this procedure is highly subjective, as the previous studies used to elicit these parameters concerned differing patient populations selected using various definitions of head injury and other entry criteria. These values were, however, discussed with the clinicians to ensure that they were in accordance with their views.

To take account of the effect of the prognostic factors on the outcome a stratified sample size calculation may be made.[6] In general, let $m$ strata $S_1, \ldots, S_m$ represent the different levels of the prognostic factors and denote the anticipated proportion of patients in $S_h$ by $s_h$, $h = 1, \ldots, m$. The parameter $\bar{p}_{h1}$ denotes the anticipated proportion of patients from both treatment groups in stratum $S_h$ having outcome category GR, while $\bar{p}_{h2}$ and $\bar{p}_{h3}$ are defined similarly for the outcomes

 

MD and SD/V/D. The total number of patients $n$ is then given by

$$n = \frac{12(u_{\alpha/2} + u_{\beta})^2}{\theta_R^2 \sum_{h=1}^{m} s_h \left(1 - \sum_{j=1}^{3} \bar{p}_{hj}^3\right)}. \quad (3)$$

Realistically, it is very difficult to obtain the information required to guess the nuisance parameters for each stratum for such a calculation to be made, and no attempt was made prior to starting the trial of eliprodil. The parameter $\theta$ now represents the log-odds ratio *within strata*. It was felt that if prognostic factors were important then the difference $\theta_R = 0{\cdot}610$ should still be detected with this re-interpretation.

It was because of the uncertainty about the proportion of patients who would fall into the three GOS categories, and about the need to stratify for prognostic factors, that a sample size review was planned. It was felt important to avoid the loss of power that can follow the use of inappropriate assumptions.

Had the study been designed sequentially, with pre-defined stopping boundaries, then the risk of underprovision of power would have been removed. The power of a sequential study is maintained, as recruitment of patients is continued until a boundary is crossed. Originally a sequential design was planned for the trial of eliprodil, using the same outcome, but at 3 months, rather than at 6. At that time, a recruitment rate of 15 patients per month was envisaged. If the study was stopped early by the sequential rule there would have been additional responses from patients already recruited, and still being followed-up, who were yet to give a 3 month outcome. According to the original plan, the number of such responses in the 'pipeline' would have been approximately 45 and thus a sequential study was thought to be feasible. However, when the 6 month outcome was adopted, and the planned recruitment rate was increased to 30 patients per month, stopping due to the sequential rule would leave approximately 180 patients in the 'pipeline'. Consequently, any advantages of sample size reduction due to early stopping would be lost. For this reason, the sequential design was abandoned.

### 2.3. Safety monitoring

Mortality was a high risk in this study and mortality and adverse event data from this trial were regularly assessed by a Data and Safety Monitoring Board (DSMB). Deaths would occur relatively quickly, and a formal sequential procedure was used to monitor death rates within 21 days of randomization. An open-top-design[18] was used, which allowed early stopping only if mortality on eliprodil substantially exceeded that of placebo and it was unethical to proceed. Otherwise, the study was to continue to its fixed sample size. The authors intend to describe this aspect of the trial of eliprodil elsewhere.

## 3. THE SAMPLE SIZE REVIEW PLAN

### 3.1. The plan

The objective of the sample size review was to preserve the 90 per cent power of the study, and thus it was intended only for checking the assumptions made in calculating the sample size, not to compare treatments. It was planned to conduct the sample size review on responses from 100 patients. Assuming an entry rate of 30 patients per month it would take just over 3 months to

 

recruit 100 patients. Six months later, just beyond 9 months into the trial, these patients would have provided their 6 month outcome scores. By that time about 180 more patients would have been recruited, which is nearly three-quarters of the initial sample size. Owing to slow initial recruitment and to dropout, it was anticipated that the sample size review was more likely to be conducted at one year.

To avoid bias, and to meet the requirements of regulatory authorities, the review was to be conducted without unblinding treatment allocation. As a result of this, the assumption of proportional odds could not be checked during the review. To check the influence of prognostic factors on 6 month GOS scores, the proportional odds model would be fitted and the likelihood ratio test used for guidance. Intervals into which prognostic factors were to be grouped were pre-specified. These groups were to be adhered to unless some of the groups had only a small number of patients, in which case groups would be combined or redefined. Factors which had a significant effect on outcome of a large magnitude would be identified and the combination of these factor groups would be used to form strata.

For each stratum, the proportion of patients in that stratum and the overall proportion of outcomes in each response category, $\bar{p}_{hj}$ would be estimated. Using these new estimates the sample size would then be recalculated using equation (3). If no prognostic factors were found to have an important effect on the outcome, then the overall proportions in each outcome category would be re-estimated and the sample size re-calculated using equation (2).

Denote the re-calculated sample size by $n_R$ and the original sample size by $n$. The new sample size actually used in the trial, $N$, was to be decided using the rule:

$$N = \begin{cases} 400 & \text{if } n_R \leqslant 400 \\ n_R & \text{if } 400 < n_R < 600 \\ 600 & \text{if } n_R \geqslant 600. \end{cases} \tag{4}$$

In this trial involving a new compound, both investigation of efficacy and safety were of primary importance, and the 6 month GOS score was not the only endpoint of interest. If a reduction in sample size based on efficacy was allowed, there may have been inadequate data to compare other responses, including safety, and so a reduction in sample size was not permitted. To give an indication of the likely range of sample size and cost for budgetary purposes, an upper limit of 600 patients (150 per cent of the initial size) was set.

After the sample size review the investigators were to be informed of the new target, and if necessary the period of recruitment was to be extended.

There was concern that the sample size review might affect the type I error rate. Gould[12] indicates that an effect is a mathematical possibility, but provides simulations for binary data demonstrating that no noticeable effect occurs. Prior to the eliprodil study simulations were performed to examine these results in the case of ordinal data, and for the extension of the sample size review to assess prognostic factors. The results are reported in the following section, and were sufficient to demonstrate that any affect on type I error rate was negligible. As a result the plan was recommended to and accepted by the trial Steering Committee.

### 3.2. Simulation results

These simulations were organized in two parts. The first part, reported in Section 3.2.1, concerned re-estimation of overall proportions in the three GOS categories, with no regard to prognostic

 

Table I. Overall and individual treatment distributions of the Glasgow Outcome Scale (GOS) under the null and alternative hypotheses as used in the simulation study, with corresponding fixed sample sizes

| Case | Hypothesis | Treatment | GOS outcome category | | | Fixed SS |
|---|---|---|---|---|---|---|
| | | | GR | MD | SD/V/D | |
| (i) | $H_0$ | either | 0·300 | 0·600 | 0·100 | 448 |
| | $H_1$ | eliprodil | 0·363 | 0·564 | 0·073 | |
| | | placebo | 0·237 | 0·636 | 0·127 | |
| (ii) | $H_0$ | either | 0·100 | 0·150 | 0·750 | 591 |
| | $H_1$ | eliprodil | 0·127 | 0·180 | 0·693 | |
| | | placebo | 0·073 | 0·120 | 0·807 | |
| (iii) | $H_0$ | either | 0·222 | 0·323 | 0·455 | 394 |
| | $H_1$ | eliprodil | 0·274 | 0·346 | 0·380 | |
| | | placebo | 0·170 | 0·300 | 0·530 | |

factors. The second part, reported in Section 3.2.2, examined the procedure of reviewing prognostic factors.

### *3.2.1. Re-estimation of outcome proportions*

The properties of the sample size review procedure were examined over three sets of overall outcome distributions. The three cases were selected to represent the following possibilities:

(i) the study patients recovering *better* than expected;
(ii) the study patients recovering *worse* than expected;
(iii) the study patients recovering *as well* as expected.

The three cases are shown in Table I, together with the corresponding fixed sample size had these rates been anticipated in advance. Case (iii) corresponds to the anticipated outcome distribution used to determine the sample size of 394. Rounding to 400 patients increases the power slightly from 0·900 to 0·905. For each of the three cases, two situations were examined giving a total of six scenarios. The first corresponds to eliprodil having no effect (under the null hypothesis) in which case the outcome proportions for each treatment were set equal to the overall outcome distribution. The second situation corresponds to eliprodil actually working (under the alternative hypothesis). Here the outcome proportions were determined from the same overall outcome proportions, under the reference improvement $\theta_R = 0{\cdot}610$. For each case the distribution of outcomes in each GOS category under the null and alternative hypotheses are also presented in Table I.

For each of the six scenarios the study was simulated 10,000 times and an initial sample size of 400 was used. For comparison, these runs were conducted both for a trial without a sample size review and for a trial with a review. Table II gives the proportion of the 10,000 trials in which the null hypothesis of no treatment effect was rejected for each scenario. Table III gives, for each case and situation, the mean and the 95th percentile of the sample size over 10,000 simulations which incorporate a review.

The 95 per cent probability interval for the significance level based on an estimate of 0·05 is (0·046, 0·054). Table II shows that the observed significance level is within this interval for the

 

Table II. Proportion of rejections of the null hypothesis in the simulation study

| Hypothesis | Case | Without review | With review |
|---|---|---|---|
| $H_0$ | (i) | 0·0482 | 0·0492 |
| | (ii) | 0·0509 | 0·0465 |
| | (iii) | 0·0470 | 0·0512 |
| $H_1$ | (i) | 0·8580 | 0·8993 |
| | (ii) | 0·7502 | 0·8921 |
| | (iii) | 0·8958 | 0·9044 |

Table III. Means and 95th percentiles of sample size in the simulation study

| Hypothesis | Case | Sample size | |
|---|---|---|---|
| | | Mean | 95th percentile |
| $H_0$ | (i) | 454 | 506 |
| | (ii) | 572 | 600 |
| | (iii) | 403 | 416 |
| $H_1$ | (i) | 454 | 506 |
| | (ii) | 572 | 600 |
| | (iii) | 403 | 415 |

procedures with or without sample size review and that the review does not materially affect the type I error rate. The 95 per cent probability interval for power based on an estimate of 0·905 is (0·899, 0·911). When there is no sample size review the power is almost achieved only if the guessed outcome distribution is correct (case (iii)), but not when the guess is wrong. Use of the sample size review procedure, however, has been largely successful in achieving the required power, irrespective of the initial guess of outcome distribution.

For case (iii), which uses the correct guesses, the estimate of power without review was 0·8958, which lies just below the probability interval. For the same case when the sample size review was conducted, the mean sample size was 403 patients, showing that in most runs no more than 400 observations were required. It is likely to be the inaccuracy of using the Normal approximation in the original sample size calculation which causes the minor deviation of power from its target setting when a sample size review was not conducted. For case (ii) the estimate of power without review was 0·7502. When a sample size review was conducted the sample size frequently had to be limited to the maximum sample size of 600, which was unsurprising as the fixed sample size calculated was 591 patients. Owing to imposition of this maximum, even when a sample size review was conducted, the estimate of power was 0·8921, which lies below the probability interval.

### *3.2.2. Assessing prognostic factors*

An examination of the effect of reviewing prognostic factors on the significance level and power of the study was made. Only the simplest case of a single factor with two levels was considered. It

 

Table IV. Proportion of rejections of the null hypothesis, mean and 95th percentile of sample size when a sample size review assessing prognostic factors was used in the simulation study

| Hypothesis | Proportion of rejections of $H_0$ | Sample size | |
|---|---|---|---|
| | | Mean | 95th percentile |
| $H_0$ | 0·049 | 528 | 600 |
| $H_1$ | 0·888 | 528 | 600 |

was assumed that the anticipated overall distribution of GOS categories at 6 months was as in Table I, case (iii), and an initial sample size of 400 was determined. Data were simulated for individual patients as follows. First a patient was allocated to stratum 1 of the prognostic factor with probability 0·4, and otherwise to stratum 2. Outcomes for patients in stratum 1 were simulated from an outcome distribution given by case (i) Table I and outcomes for patients in stratum 2 were simulated from the outcome distribution given by case (ii), both under the null (situation 1) and alternative hypothesis (situation 2). Thus, for both situations there was a clear stratum effect.

At the review stage the influence of the prognostic factor was assessed without breaking the blind. First a $3 \times 2$ contingency table of the GOS category against the stratum was formed. The significance of the relationship between these two was determined using a chi-square test of independence at the 5 per cent level of significance. If the effect of the prognostic factor was significant, then the sample size $n_R$ was re-calculated using equation (3). To perform this calculation the proportion of patients $s_h$ in each stratum and the proportions of those having outcome category GR, MD or SD/V/D $(\bar{p}_{h1}, \bar{p}_{h2}, \bar{p}_{h3})$ were estimated from the initial sample of 100. The new sample size was set to $n_R$, except that lower and upper bounds of 400 and 600 were imposed according to the rule given by (4). The simulation of patients' data was then continued to that new sample size. In practice the decision to stratify should not be based solely on achieving a 5 per cent significance level, and furthermore a more sophisticated assessment based on likelihood ratio testing could be used. For the purpose of simulation, a method which was both automatic and quick to compute was needed.

In the final analysis, the test of treatment effect was adjusted for influential prognostic factors using stratification, and methods devised for meta-analysis[19] were used to test the homogeneity of treatment effects over the two strata. Under both the null and alternative hypothesis, 10,000 simulations were performed. In all of these simulations the test for stratum effect conducted at the sample size review was significant. Table IV gives the proportion of rejections of the null hypothesis of no significant treatment effect, the mean and 95th percentile of sample size for each situation. Thus the target type I error rate was achieved, and the target power of 0·905 was well approximated.

Further simulations for the two situations above were made in which the sample size review did not include an examination of an adjustment for influential prognostic factors. These gave a type I error rate of 0·049, but a power of only 0·810. These results indicate the importance of checking for influential prognostic factors and adjusting the sample size accordingly at the review stage.

 

*3.2.3. Summary*

The simulation study demonstrated two results. First, that the sample size review does not affect the type I error rate, and second that the review procedure is successful in achieving the intended power. The simulations also illustrated that not using a review can result in an appreciable loss of power, unless nuisance parameters are guessed correctly. The importance of adjusting the sample size at the review for influential factors to achieve the intended power was also demonstrated.

## 4. IMPLEMENTATION

### 4.1. Data available

Data from 93 patients were supplied by the sponsor for the sample size review. Recruitment to the study started on 22 February 1993, and, after slower recruitment than had been planned, the 93rd patient was recruited a year later on 22 February 1994. The review was undertaken in February 1995, once the necessary data had been collected and validated. Of the 93 patients, 86 provided a 6 month GOS score. For 6 patients who were alive with no 6 month GOS score, their last GOS evaluation was used. One patient was excluded as no GOS score at all was available. Therefore 92 patients were used for the sample size review.

To examine the influence of the prognostic factors, the following stratification levels were used: GCS (4–5, 6–8), delay before administration of treatment ($0 < 10,\ \geqslant 10$ hours) and age (16–25, 26–60 years).

### 4.2. Recalculation of sample size

The only factor found to have significant effect of large magnitude in a proportional odds regression analysis of the 6 month GOS was GCS at baseline ($p = 0{\cdot}0009$); none of the other factors were significant. Patients with a GCS greater than 5 had a higher odds of a better recovery. As a result the GCS was used to form a stratification of these data. The proportion of patients in each stratum and overall (pooled over treatments) proportions of the three GOS scores within each stratum were estimated (see Table V).

Using the values of Table V, the sample size was recalculated using equation (3). The sample size required is proportional to the inverse of the conversion factor $(1 - \sum_{j=1}^{3} \bar{p}_j^3)$ estimated from the nuisance parameters. The value used in the initial sample size was 0·8608. At the review it was 0·7676 for the (GCS $\leqslant$ 5) stratum and 0·7616 for the (GCS $>$ 5) stratum. Although the conversion factors are similar between the two strata there is a considerable difference between strata in the proportions of each of the three GOS scores and from the distribution of GOS scores originally guessed. Averaging these conversion factors, weighting by the proportion in each stratum, gives a conversion factor of 0·7640. The ratio of the original conversion factor to this new value is $0{\cdot}8608/0{\cdot}7640 = 1{\cdot}127$ which, when applied as a multiplier to the original sample size of 394, determined that the new sample size required was 444 patients, which was rounded to 450 patients.

For comparison, Table VI shows the proportion of each of the GOS outcome scores as originally anticipated, and as would arise from an unstratified sample size review. These show considerable differences, particularly in the GR and MD categories. Without adjustment for prognostic factors the sample size would have been 409 patients.

 

Table V. GOS distribution stratified for GCS at the sample size review of the study of eliprodil

| GCS at day 0 | Proportion of patients | GOS category | | |
|---|---|---|---|---|
| | | GR | MD | SD/V/D |
| $\leqslant 5$ | 0·402 | 0·270 | 0·135 | 0·595 |
| $> 5$ | 0·598 | 0·600 | 0·127 | 0·273 |

Table VI. Overall GOS distribution anticipated and re-estimated (unstratified) at the sample size review of the trial of eliprodil

| Source | GOS category | | |
|---|---|---|---|
| | GR | MD | SD/V/D |
| Anticipated | 0·222 | 0·323 | 0·455 |
| Sample size review | 0·467 | 0·131 | 0·402 |

The assumption of proportional odds was accepted by the Steering Committee for this trial. As the review was conducted without unblinding treatment allocation the assumption of proportional odds could not be checked. However, on completion of the study various goodness-of-fit techniques were applied. These showed that the assumption, whilst not fully valid, was not misleading. Full results of these checks are not reported here due to continuing confidentiality of the details of the final trial results.

## 5. DISCUSSION

In this paper, simulation evidence has been presented which demonstrates that for a clinical trial with ordered categorical responses, a sample size review is effective in preserving the intended power, while having a negligible effect on type I error. The implementation of the method in a clinical trial in head injury has been described; the recommendation of modest increase in sample size from 400 to 450 patients was accepted by the trial Steering Committee without hesitation.

Guesses of the nuisance parameters required for sample size calculation are not easy to obtain. The values used may be based on similar studies conducted previously. Even if trials of similar treatments do exist there will often be other variations such as entry criteria including less serious cases or different timings of recording the outcome variable. Often no similar trials will have been conducted previously, apart from those in early phases. The small size of these trials will prohibit reliable estimates of the outcome distribution to be made, particularly for estimates stratified for prognostic factors. In addition early phase trials may well have recorded a different endpoint. Using data from publications involves the same problems, to a greater degree. There is therefore much scope for use of sample size reviews to remove the guessing element of sample size calculations.

Approaches similar to that described here can be made in clinical trials for which the primary response is neither binary nor ordinal. The application of sample size reviews to Normally

 

distributed outcomes has been discussed by Gould and Shih.[20] The objective in that case is a re-evaluation of the standard deviation of responses. This can be achieved, even from blinded data, using the EM algorithm.

For survival data a special problem arises. Pre-trial translations from the number of events required to the number of patients to achieve a given power involve anticipated values of the overall survival function $\bar{S}(t)$ of patients in the trial. Values are required for times $t$ between 0 and the intended maximum follow-up time $t_f$. Denote the anticipated values by $\bar{S}_a(t_i)$, $i = 1, \ldots, f$. At the sample size review $\bar{S}(t_i)$ can be reassessed only for values of $t_i$ less than the current duration of the trial: $t_1, \ldots, t_c$, $c < f$, say. The new estimates can be denoted by $\bar{S}_{new}(t_i)$, $i = 1, \ldots, c$. A reasonable method of completing the revised overall survival function is as follows. Let $\phi$ denote the average difference between the anticipated and new values of $\bar{S}(t_i)$ on the complementary log-log scale for $i = 1, \ldots, c$:

$$\phi = \frac{1}{c} \sum_{i=1}^{c} [-\log\{-\log \bar{S}_{new}(t_i)\} + \log\{-\log \bar{S}_a(t_i)\}].$$

Then find values of $\bar{S}_{new}(t_i)$ for values of $t_i$ beyond the current duration of the trial from

$$-\log\{-\log \bar{S}_{new}(t_i)\} = -\log\{-\log \bar{S}_a(t_i)\} + \phi. \qquad i = c+1, \ldots, f.$$

The requirement that a sample size review be conducted blind is a defence against the temptation to use the treatment labels in an interim assessment of relative efficacy. It also helps to avoid the suspicion that such an assessment was conducted, even if it was not. The only situation, amongst those considered in this section, in which blindness leads to a specific methodological complication is that of Normally distributed data. In that case conventional estimates of the unknown common variance require unblinding and alternative methods are needed.[20] However, blindness does rule out the checking of model assumptions such as proportional odds or proportional hazards which also affect sample size. Additional considerations which can be added to a sample size review are assessments or reassessments of rates of withdrawal and protocol violation; corrections to sample size for these can then be reworked.

A trial design with a sample size review is in some senses a compromise between a fixed sample size design and a sequential plan. Sequential designs involving plotting against information rather than sample size,[21] guarantee power. The effect of poor guesses of nuisance parameters is to render predictions of expected or median sample size inaccurate. The sequential design, by including unblinded interim treatment comparisons, of course allows early stopping as soon as the trial conclusion is clear.

ACKNOWLEDGEMENTS

The authors are grateful to Synthelabo Recherche for permission to publish these data, and in particular would like to thank Andreas Zipfel, Christian Giroux, Christiane L'Héritier and Bernard Sebastien for their collaboration.

## REFERENCES

1. Machin, D., Campbell, M. J., Fayers, P. M. and Pinol, A. P. Y. *Sample Size Tables For Clinical Studies*, 2nd edn, Blackwell Science, Oxford, 1997.
2. Pocock, S. J. *Clinical Trials: A Practical Approach*, Wiley, Chichester, 1983.
3. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn, Wiley, New York, 1981.

 

4. Freedman, L. S. 'Tables of the numbers of patients required in clinical trials using the logrank test', *Statistics in Medicine*, **1**, 121–129 (1982).
5. Collett, D. *Modelling Survival Data in Medical Research*, Chapman and Hall, London, 1994.
6. Whitehead, J. 'Sample size calculations for ordered categorical data', *Statistics in Medicine*, **12**, 2257–2271 (1993).
7. Hilton, J. and Mehta, C. R. 'Power and sample size calculations for exact conditional tests with ordered categorical data', *Biometrics*, **49**, 609–616 (1993).
8. Lesaffre, E., Scheys, I., Fröhlich, J. and Bluhmki, E. 'Calculation of power and sample size with bounded outcome scores', *Statistics in Medicine*, **12**, 1063–1078 (1993).
9. Kolassa, J. E. 'A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data', *Statistics in Medicine*, **14**, 1577–1581 (1995).
10. Hilton, J. F. 'The appropriateness of the Wilcoxon test in ordinal data', *Statistics in Medicine*, **15**, 631–645 (1996).
11. Wittes, J. and Brittain, E. 'The role of internal pilot studies in increasing the efficiency of clinical trials', *Statistics in Medicine*, **9**, 65–72 (1990).
12. Gould, A. L. 'Interim analyses for monitoring clinical trials that do not materially affect the type I error rate', *Statistics in Medicine*, **11**, 55–66 (1992).
13. Gould, A. L. 'Planning and revising the sample size for a trial', *Statistics in Medicine*, **14**, 1039–1051 (1995).
14. Robinson, L. D. and Jewell, N. P. 'Some surprising results about covariate adjustment in logistic regression models', *International Statistical Review*, **59**, 227–240 (1991).
15. Ford, I., Norrie, J. and Ahmadi, S. 'Model inconsistency, illustrated by the Cox proportional hazards model', *Statistics in Medicine*, **14**, 735–746 (1995).
16. Scatton, B., Giroux, C., Thenot, J. P., Frost, J., George, P., Carter, C. and Benavides, J. 'Eliprodil Hydrochloride', *Drugs of the Future*, **19**, 905–909 (1994).
17. McCullagh, P. 'Regression models for ordinal data', *Journal of the Royal Statistical Society*, *Series B*, **43**, 109–142 (1980).
18. Whitehead, J. 'Sequential methods based on boundaries approach for the clinical comparison of survival times', *Statistics in Medicine*, **13**, 1357–1368 (1994).
19. Dersimonian, R. and Laird, N. 'Meta-analysis in clinical trials', *Controlled Clinical Trials*, **7**, 177–188 (1986).
20. Gould, A. L. and Shih, W. J. 'Sample size re-estimation without unblinding for normally distributed data with unknown variance', *Communications in Statistics – Theory and Methods*, **21**, 2833–2853 (1992).
21. Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, revised 2nd edn, Wiley, Chichester, 1997.

 