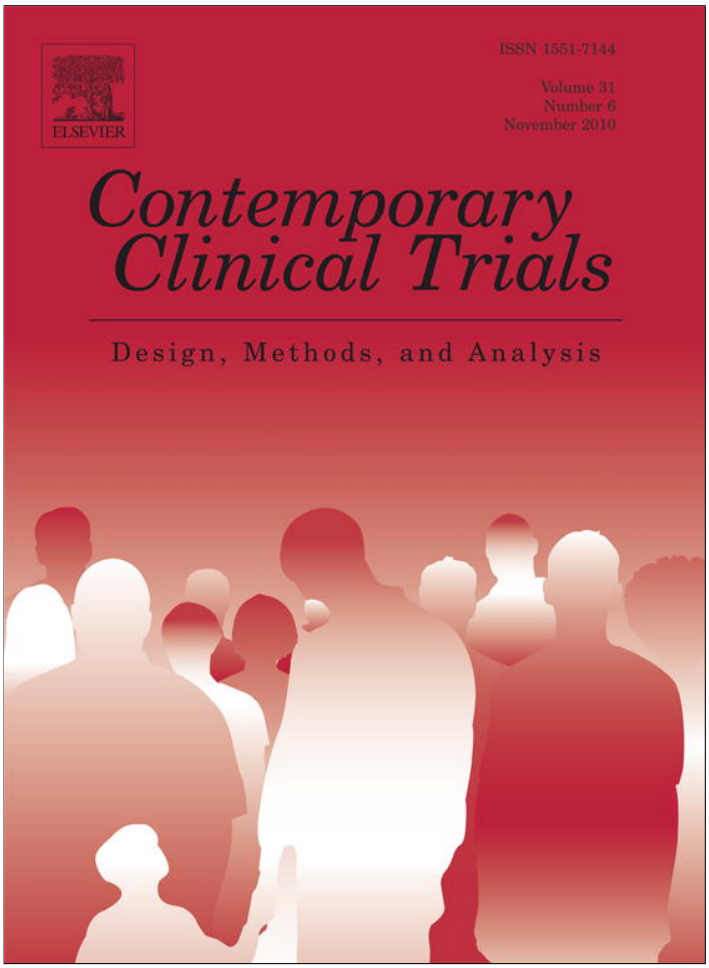

Contemporary Clinical Trials 31 (2010) 579–586

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

The use of mid-trial reviews for design modifications in small scale clinical studies

M.R. Sooriyarachchi [a,*], R.V. Jayatillake [a], H. Ranganath [b], Michael Eddleston [c,d]

[a] Department of Statistics, University of Colombo, Sri Lanka
[b] South Asian Clinical Toxicology Research Collaboration, Sri Lanka
[c] Clinical Pharmacological Unit, University of Edinburgh, UK
[d] National Poisons Information Service - Edinburgh, Royal Infirmary Edinburgh, Edinburgh, UK

ARTICLE INFO

Article history:
Received 12 May 2010
Accepted 22 July 2010

Keywords:
Internal pilot study
Mid-trial design review
Normal error linear model
Sample size reestimation
Testing distributional and modelling assumptions

ABSTRACT

Many clinical studies such as those in the areas of toxicology, early phase clinical trials and bioequivalence studies use small samples due to the high cost of experimentation. These studies test hypotheses based on small samples. These small samples result in low power and therefore even if the alternative hypotheses may be true the chance of it being rejected is low. The sample size is determined in an ad-hoc way and no proper scientific approach is used. Sample size calculations for clinical studies are usually conducted to determine the total number of patients needed to satisfy a specified power requirement, and their validity is dependent on pre-trial knowledge of nuisance parameters and distributional and modelling assumptions. Another short coming is that often hypotheses are tested without checking the assumptions required by the test. This paper looks at design reviews in the context of small samples. It examines several design modifications done with asmall internal pilot study. In the past similar techniques have been applied to large scale studies but it sperformance is yet to be established in small scale clinical studies thus the contribution of this paper is in justifying the validity of these techniques for small samples too. The methodology is illustrated on an uncontrolled observational toxicology study. In this paper simulations will be presented showing that the design modifications would not influence the type-I error rate and that these would be successful in preserving the power, and the implementation of the design review procedure will be described.

1. Introduction

Wittes and Britain [1], Gould [2,3] and Gould and Shih [4] developed the idea of mid-trial reviews using internal pilot studies for the purpose of estimating unknown parameters required to determine sample size. More recently Zucker et al. [5], Shih [6], Gould [7], Whitehead et al. [8] and Bolland et al. [9] among others discuss the applicability of these mid-trial reviews for making other decisions in addition to estimating unknown parameters. Coffey and Muller [10] use an internal pilot study for fitting a Normal error linear model and estimating the variance required to determine the sample size using the mean square error (MSE) of the model. Friede and Kieser [11] review past studies on this topic. While all these authors have considered the performance of design reviews in large scale clinical trials its validity in small scale clinical studies is yet to be established [12]. The objective of this paper is to examine the performance of several design modifications at the design review, based on a small scale clinical study in order to determine the validity of these methods in small studies.

The methods are illustrated on an uncontrolled observational study of the drug atropine in the treatment of

* Corresponding author. Department of Statistics, University of Colombo, P.O. Box 1490, Colombo 3, Sri Lanka. Tel.: +94 2590111, +94 5731269; fax: +94 2587239.
E-mail address: roshini@mail.cmb.ac.lk (M.R. Sooriyarachchi).

1551-7144/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2010.07.011

bradycardia (slow pulse) in yellow oleander seed poisoned patients [13,14]. The design modifications applied to this study are sample size recalculation by estimating nuisance parameters and testing of pre-trial distributional and modelling assumptions through a mid-trial review. Prior to completing the review a simulation study is performed to confirm that the review procedure is successful in preserving power without inflation of the type-I error rate, and those simulation results are presented here.

The primary response in the study is the change in heart rate 5 min after treatment and baseline. Medically, there were two objectives of equal interest in this study, the first was to determine whether 0.6 mg of atropine significantly increases the heart rate at 5 min after treatment from that at baseline. The second was to determine whether a measure of the rise in heart rate (W) is related to a measure of the heart rate at baseline (Z). These two objectives require the calculation of two sample sizes and selecting the larger of the two to satisfy both power requirements.

In this paper the sample size formulae required for satisfying the two objectives of the study are derived. These sample size formulae involve nuisance parameters for which values were unknown when the trial was set up. For example, with normally distributed data the variance of the primary response variable is required. In addition knowledge of the form of the most suitable model between a measure of the change in heart rate (W) and a measure of the heart rate at baseline (Z) together with determination of W and Z is needed. This means that it is required to identify a suitable relationship between a measure of the difference in heart rate (W) and a measure of baseline heart rate (Z). The form and the error structure of such a model are unknown at the design stage.

2. Materials and methods

2.1. Study description

The yellow oleander tree is found commonly in Sri Lanka where its seeds have become a popular means of self harm [13,14]. A common consequence of poisoning is a bradycardic (slow) pulse [15]. Atropine is the recommended treatment for oleander-induced bradycardia. It is common practice to give small doses of Atropine when the heart rate is above 60 bpm with the aim of keeping the heart rate around 80 bpm. Doses as large as 12 mg over 1 h have been used for treatment in hospitals in Sri Lanka [16]. However, bradycardia has been effectively treated with 0.6 mg of atropine in uncontrolled studies and large doses of atropine can result in confusion and hyperpyrexia [14]. An uncontrolled observational study was undertaken in a Sri Lankan hospital in the North Central province to determine the typical heart rate response to 0.6 mg of atropine. Eligible patients were those with yellow oleander seed self-poisoning whose heart rate was below 90 bpm at baseline. Usually, Bradycardia is defined as a low frequency heart rate below 60 bpm. In our study the criteria for selection was below 90 bpm because in Yellow Oleander poisoned patients the heart rate and Bradycardia status is very variable and could vary among normal, mild, moderate and severe status within a short period of time. This is the reason that even patients having heart rate between 60 and 90 bpm are included in this study as a precaution that their heart rate would not go below 40 bpm [17,18]. Patients treated at local hospitals prior to the admission to the study hospital and patients given more than 0.6 mg of atropine were excluded from this study. Atropine was given intravenously to each patient as part of their routine clinical care. Patients were then observed for 60 min and information on the primary response variable, heart rate, was collected 5 times at baseline, 2, 5, 15 and 60 min after the treatment was given. Medically it is known that the peak heart rate after atropine injection in anaesthetized patients undergoing surgery occurs at 5 min [19]. In Yellow Oleander poisoned patients some work on this has been done by Eddleston (2007) [20] and he has found the same result. Thus the medical interest lay only in the time to maximum heart rate and no Statistical adjustments were made for repeated measures. This is further justified by the fact that, our atropine study which was continued for 66 patients with Yellow Oleander poisoning found that on average the peak heart rate occurs at 5 min. Thus of these time points heart rate at 5 min was selected to be compared to heart rate at baseline because it is known that the maximum heart rate after atropine administration on average occurs at 5 min. The primary outcome variable, heart rate, is a continuous measure. An important prognostic factor was thought to be the baseline heart rate. Ethics approval for observational studies of poisoned patients was obtained from Oxford Tropical Medicine Research Ethics Committee and the Colombo Faculty of Medicine Ethics Committee. Patient consent was not obtained for this observational study of response to usual clinical care, as no interventions were given as part of the trial.

2.2. Sample size calculation

2.2.1. Setting up of notation

A standard form of power requirement for the first objective specifies that a significant increase in heart rate at 5 min from that at baseline should be found with probability $(1-\beta_1)$ if a true increase equal to θ_R is present. By 'finding a significant increase' means obtaining a significant difference at the 100α percent level against the upper one sided alternative. The probability $(1-\beta_1)$ is referred to as the power of the test for satisfying the first objective.

The value θ_R represents a clinically relevant increase in heart rate and is a particular value of some measure θ of change in heart rate. Sample size formulae and tables for controlled trials aiming to examine the familiar hypothesis of efficacy, for a variety of types of end point are given by Machin et al [21].

A power requirement for the second objective specifies that a significant correlation between W and Z should be detected with power $(1-\beta_2)$ if a true value of correlation R (<0) is present. This corresponds to a one sided alternative. The notation can be interpreted similarly as for the first objective. As there are two tests here the Bonferroni correction is used for multiple comparisons and therefore a more stringent significance level of α is used for both tests so as to keep the overall significance level at $2\alpha = 5\%$.

As there were two questions of interest, two sample sizes had to be calculated. The first sample size was based on the

requirement to detect an increase of 10 bpm in mean heart rate between 5 min after treatment and baseline with 80% power at the 2.5% significance level. The second sample size was based on the requirement to detect a correlation of −0.5 between a measure of the rise in heart rate at 5 min (W) and a measure of the baseline heart rate (Z) with 80% power at the 2.5% significance level. The significance level was taken to be $\alpha = 2.5\%$ so that the overall significance level, $2\alpha = 5\%$. As the relationship of interest between W and Z was a linear one, correlation was used in preference to regression in calculating the second sample size. Both sample sizes calculated are approximate and not exact.

2.2.2. Calculation of the first sample size

As mentioned previously W is the measure of change in heart rate which is of interest. Let μ_w be its true mean. Then the hypothesis to be tested is that of no change in the true mean heart rate at 5 min and baseline (H_{01}:$\mu_w = 0$) versus the alternative that there is an increase in the true mean heart rates at 5 min and baseline (H_{11}:$\mu_w = \theta_R > 0$). Here θ_R is the difference in mean heart rate between 5 min and baseline which requires to be detected as significant if a difference exists. It is assumed that W follows an approximate normal distribution. This assumption will be tested at the review stage. Using the sample size formula of Pocock [22] the total number of patients required for satisfying the power requirement $(1-\beta_1)$ for testing H_{01} is n_1, where,

$$n_1 = \frac{\left[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta_1)\right]^2 \sigma^2}{\theta_R^2} \tag{1}$$

Here σ^2 denotes the variance of W, $\Phi^{-1}(u)$ is the u^{th} percentile of the standard normal distribution, α corresponds to the significance level and $(1-\beta_1)$ corresponds to the power.

In order to calculate this, the form of a suitable Gaussian error linear model between W and Z and the mean square error (MSE) is needed to estimate σ^2 but is unknown at the planning stage. Since there is only one treatment group in this study blinding is not an issue. However it might be if there was more than one treatment. In this case the EM estimator (based on the EM algorithm) [2,4] is a possible alternative for preserving the blindness, though in recent times its use has been criticized for small interim samples as being biased and inefficient, compared to the maximum likelihood estimator [23]. A modified EM estimator [23] has been found to be more effective for blind estimation in small samples, though more simulation studies are needed to confirm its performance.

2.2.3. Calculation of the second sample size

Suppose, the population correlation coefficient between W and Z is denoted by R and its estimate based on the sample data is denoted by r. Then the hypothesis to be tested is that there is no correlation between W and Z (H_{02}:$R = 0$) versus the alternative that there exists a negative relationship between W and Z (H_{12}:$R = R' < 0$). The general consensus in medical circles is that small baseline values of heart rate are associated with large changes and large baseline values are associated with smaller changes. Thus a negative relationship is anticipated between W and Z. Here R' is the value of correlation required to be detected as significant if such a correlation exists.

Using Fisher's Z transformation [24] the test statistic $Z = \frac{1}{2} log\left(\frac{1+r}{1-r}\right)$ is such that Z follows a normal distribution given by

$$Z \sim N\left(\frac{1}{2} log\left(\frac{1+R}{1-R}\right), \frac{1}{n-3}\right) \tag{2}$$

where n is the sample size required.

Using the power approach of Pocock [22] and this test statistic, the second sample size formula is given by,

$$n_2 = \frac{\left[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta_2)\right]^2}{\left[\frac{1}{2} log\left(\frac{1+R'}{1-R'}\right)\right]^2} + 3 \tag{3}$$

For values of $\alpha = 2.5\%$, $1-\beta_2 = 80\%$, $R' = -0.5$ the second sample size n_2 results in 29 observations.

2.3. The plan for the internal pilot study

2.3.1. The plan

The method of Wittes and Brittain [1] was used to do an internal pilot study at n_2 (29) observations, to identify a suitable model and to obtain the MSE of this model. Here n_2 is the sample size required to satisfy power requirement II. In obtaining a suitable model, the following strategy was used. Two medically plausible variables were considered for W that is a measure of the change in heart rate at 5 min and baseline. These two measures were the absolute difference of the two measures (diff) and the logarithm of the rate of the two measures (lrate). Similarly two measures of the baseline heart rate, Z, were also used. This being the direct baseline heart rate (bhr) and its reciprocal (rbh). Four correlation coefficients were estimated for this 2×2 matrix of variables, using the first 29 observations. The preferred pair of variables was the simplest that is diff and bhr. However, the correlation of the most significant pair of variables was compared with the correlation of the preferred pair of variables using a rule of thumb. This rule of thumb was to use the pair of variables with the highest correlation coefficient if its correlation coefficient was at least greater than 0.2 on an absolute scale from that of the preferred pair of variables. Based on this either the preferred model or if a more appropriate model exists, this new model was selected. In the past (before 2000) Statistics used in the analysis had to be those used in the Statistical design. However with the advent of Adaptive methods [25], there is no longer this restriction.

The method of Coffey and Muller [10] was used to calculate the first sample size,

$$n_1 = \frac{MSE\left(\Phi^{-1}(\alpha) + \Phi^{-1}(\beta_1)\right)}{\theta_R^2}. \tag{4}$$

The new sample size actually used in the trial, n, was decided using the rule

$$n = \begin{cases} n_2 & \text{if } n_2 \geq n_1 \\ n_1 & \text{if } n_1 \geq n_2 \end{cases}. \tag{5}$$

2.3.2. Testing hypothesis at the end of the study

2.3.2.1. The first hypothesis. Suppose the Gaussian error linear model selected is $W_i = \beta_0 + \beta_1 Z_i + \xi_i$ where W_i is a measure of the rise in heart rate and Z_i is a measure of the baseline heart rate of the ith patient. The notations β_0, β_1, and ξ_i are the familiar notations used in linear regression [26]. The test statistic t_0 for testing the null hypothesis H_{01} is such that $t_0 = \frac{\overline{w}}{\sqrt{MSE_1 / n}}$, where $\overline{w}$ is the sample mean of W based on the total observations, MSE_1 is the mean square error of the model between W and Z based on only the first stage observations and n is the new sample size. The test statistic t_0 follows a student's t distribution with n_2-2 degrees of freedom. This statistic is used to test the first hypothesis. As the model selection is done at the interim stage taking the MSE of the total observations for calculating the test statistic inflates the type I error. The method of Stein [27], used here overcomes this problem by using the MSE based only on the first stage observations.

2.3.2.2. The second hypothesis. Fisher's Z [24] statistic is only approximately normal. The normality improves for large n. In order to make the statistic more normal, Fisher [24] introduced a bias correction. The bias corrected test statistic $z'' = z' - \frac{r}{2(n-1)}$ which follows an approximate normal distribution given by $z''\sim N\left(tanh^{-1}R, \frac{1}{n-3}\right)$ is used as the test statistic here.

3. Results

3.1. First implementation of plan

The results from analyzing the first 29 observations are presented in this section. The first objective here is to identify a suitable Gaussian error linear model between a measure of the rise in heart rate (W) and a measure of the baseline heart rate (Z). Several such measures are considered and correlations of such measures are given in Table 1.

The correlations show that the strongest linear relationship is between the log rate of heart beat at 5 min and baseline (lrate) and the reciprocal of heart rate (rbh). However, this is not very different from that between the absolute difference of heart rate at 5 min and baseline (diff)

Table 1
Correlations of several measures of rise in heart rate and baseline heart rate for the first 29 observations. The values within brackets correspond to p-values associated with the correlation.

Measure of rise in heart rate	Measure of baseline heart rate	
	Bhr	rbh
diff	-0.461 (0.0118)	0.468 (0.0105)
Lrate	-0.572 (0.0012)	0.584 (0.009)

Where,
diff is the absolute difference in the heart rate between 5 min and baseline.
lrate is the log of the ratio of heart rate at 5 min and heart rate at baseline.
bhr is the baseline heart rate.
rbh is the reciprocal of bhr.

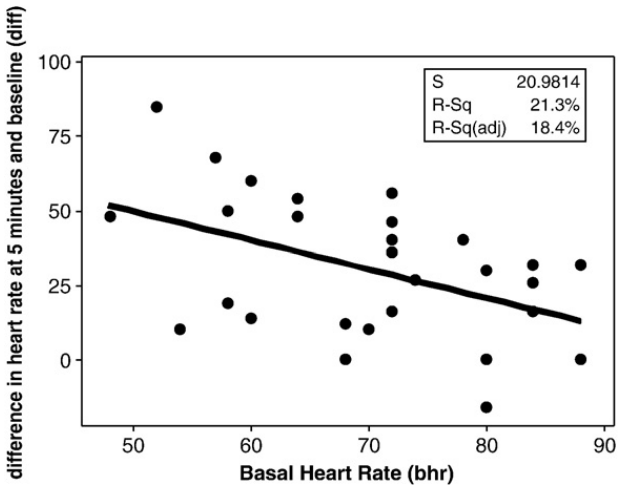

Fig. 1. Regression plot of difference in heart rate at 5 min after treatment and at baseline (diff) and baseline heart rate (bhr) with its corresponding Gaussian error linear model incorporated based on the first 29 observations.

and the baseline heart rate (bhr). Thus W is taken as diff and Z is taken as bhr. Fig. 1 illustrates this relationship. This figure shows a scatter plot of diff and bhr with its corresponding Gaussian error linear model incorporated. Fig. 1 indicates that there is an approximate linear negative relationship between diff and bhr.

Fig. 2 shows a normal probability plot of the errors with 95% confidence intervals. Anderson Darling test [28] and confidence intervals show that assumption of normality is approximately satisfied. The F test in the Analysis of Variance (ANOVA) is fairly robust to departures from normality. Thus unless there is extreme departures from normality, in which case other transformations should be used, there is no cause for alarm.

Fig. 3 shows a plot of the residuals versus fitted values. There is no noticeable pattern and the residuals are distributed in a band around zero, within 95% confidence limits ± 2. (If the standardized residuals are N(0,1) then the limits will be $\Phi^{-1}(0.975) \cong 2$ where Φ^{-1} is the inverse cumulative distribution function of the N(0,1)).

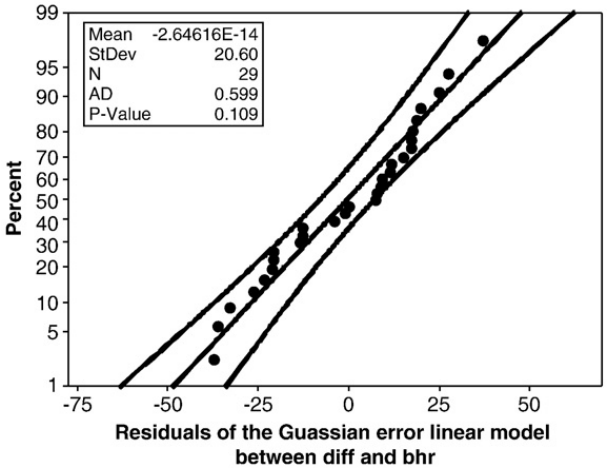

Fig. 2. Normal probability plot of the errors of the Gaussian error linear model between the difference in heart rate at 5 min after treatment and at baseline (diff) and baseline heart rate (bhr) fitted to the first 29 observations.

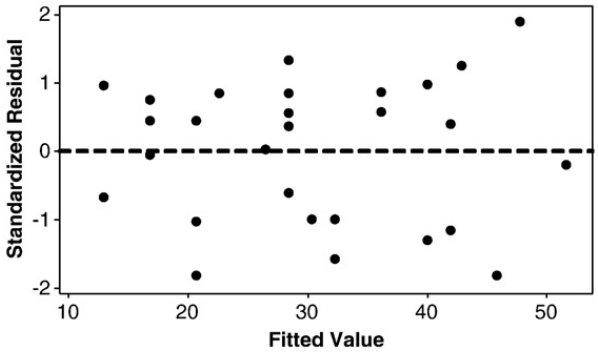

Fig. 3. Plot of the residuals versus fitted values of the Gaussian error linear model between the difference in heart rate at 5 min after treatment and at baseline (diff) and baseline heart rate (bhr) fitted to the first 29 observations.

Table 2
ANOVA for the model between difference in heart rate at 5 min after treatment and at baseline (diff) and baseline heart rate (bhr) fitted to the first 29 observations.

Source of variation	Degrees of Freedom	Sum of squares	Mean sum of squares	F-value	p-value
Due to Regression	1	3212.3	3212.3	7.3	0.012
Due to Residual Error	27	11,885.9	**440.2**		
Total Variation	28	15,098.1			

The Analysis of variance (ANOVA) for this chosen model is given in Table 2.

Table 2 shows that there is a significant linear relationship between difference in heart rate and basal heart rate. The MSE of this model is 440.2. This will be used as an estimate of σ^2 for estimating the first sample size, n_1.

3.2. Simulation study

3.2.1. Parameters used

Some practically plausible values were taken for the parameters required for the simulation study. A mean heart rate at baseline (μ_1) of 70 bpm was used. The time of interim inspection was decided to be at 29 patients. Five values of reference improvement $\theta_R = 5,7.5,10,20,30$ were examined. For the Mean of baseline heart rate (Z), variance of baseline heart rate (Z) and variance of errors (variance of diff (W)) values based on the data collected, that is on the 29 observations were used. These values are 70.103, 122.88 and 440.22 respectively. The reference correlation coefficient between W and Z (R') was taken to be -0.5.

The properties of the mid-trial review were examined over four scenarios. These four scenarios represent the following four hypotheses:

Null hypothesis H_{01}, Null hypothesis H_{02} (that is taking $\theta = 0$ and $R = 0$)

Null hypothesis H_{01}, Alternative hypothesis H_{12} (that is taking $\theta = 0$ and $R = R' < 0$)

Alternative hypothesis H_{11}, Null hypothesis H_{02} (that is taking $\theta = \theta_R > 0$ and $R = 0$)

Alternative hypothesis H_{11}, Alternative hypothesis H_{12} (that is taking $\theta = \theta_R > 0$ and $R = R' < 0$).

A 10,000 simulations of the model $diff = \beta_0 + \beta_1(bhr) + \xi_i$ were carried out for each of the four scenarios up to the interim stage. In the simulated model β_0 is the intercept, β_1 is the slope and ξ is the error term.

At the interim stage from the model fitted alternative values for W and Z such that $W = ln\left(\frac{diff + bhr}{bhr}\right) = lrate$ and $Z = \frac{1}{bhr} = rbh$ were derived and the 2×2 correlation matrix of the values of W and Z determined. The value of W was taken to be diff and Z taken to be bhr if the correlation coefficient between diff and bhr was not less than 0.2 (on the absolute scale) than the correlation coefficient between lrate and rbh. Otherwise W was taken to be lrate and Z was taken to be rbh. Thus the model could be switched at the interim stage. The other two correlations in the correlation matrix were not considered in the simulations as it would have complicated matters. However this procedure could be generalized for all correlations, the principle being the same. After the model was selected the remaining data was simulated under the selected model and the hypotheses tests conducted. Both tests were considered to be one sided. Here α was taken to be 2.5% for both tests and $1 - \beta$ was taken to be 80% for both tests.

3.2.2. Results of the simulation study

Table 3 gives the proportion of rejections of the two null hypotheses and the number of times out of 10,000 that the model was changed, for the four scenarios for the different parameter combinations when a mid-trial review was used in the simulation study.

Table 4 gives the average sample size for the four scenarios for the different parameter combinations when a mid-trial review was used in the simulation study.

3.2.3. Summary of the simulation study

The 95% probability interval for the significance level based on a sample of 10,000 and estimate of 0.025 is (0.022, 0.028). The results in Table 3 show that over all combinations of θ_R examined, the observed significance level is within this interval for both tests and thus the review does not materially affect the type I error rate. The 95% probability interval for power based on an estimate of 0.80 is (0.792, 0.808). For large θ_R, that is when n_2 is greater than n_1 the power for the second test is well maintained but as expected the first test is over-powered. Similarly when θ_R is small, that is when n_1 is greater than n_2 power for the first test is well maintained but the second test is over-powered. Use of the review procedure has been successful in achieving the required power for both tests. The results in table 3 also show that changing the model at interim stage has not affected the type I error materially.

The results in Table 4 show that when θ_R is small (n_1 is greater than n_2) then the sample sizes are quite large and varies between the scenarios as the sample size will then depend on n_1 which in turn depends on MSE which is variable. On the other hand when θ_R is large n_2 is greater than n_1 and this is fixed.

The model selection procedure has not affected the error rates. This is because Stein's procedure [22] has been used in the calculation of the t-test statistic. Further simulations (not

Table 3
Proportion of rejections of the null hypothesis under various scenarios when a mid-trial review was used in the simulation study. This table also shows the number of times out of 10,000 that the model W = lrate and Z = rbh was used instead of W = diff and Z = bhr.

θ_R	Scenario											
	1			2			3			4		
	H_{01}	H_{02}	No: of times (out of 10000) model change	H_{01}	H_{12}	No: of times (out 10000) model change	H_{11}	H_{02}	No: of times (out of 10000) model change	H_{11}	H_{12}	No: of times (out of 10000) model change
5	0.026	0.026	135	0.025	1	0	0.804	0.025	59	0.798	1	0
7.5	0.024	0.026	150	0.025	0.992	0	0.813	0.025	45	0.798	0.991	0
10	0.025	0.024	129	0.028	0.921	2	0.811	0.025	34	0.819	0.923	3
20	0.024	0.025	116	0.023	0.829	1	0.999	0.023	489	0.999	0.831	55
30	0.023	0.024	146	0.023	0.833	0	1	0.021	2456	1	0.819	442

Table 4
Average new sample size under various scenarios when a mid-trial review was used in the simulation study.

θ_R	Scenario			
	1	2	3	4
5.0	153	150	152	149
7.5	69	67	67	67
10	40	38	39	39
20	29	29	29	29
30	29	29	29	29

reported here) show that the usual *t*-test statistic inflates the type I error.

3.3. Second implementation of plan

3.3.1. Recalculation of sample size 1 (n_1)

From Eq. (4), n_1 is determined by $\frac{MSE(\Phi^{-1}(\alpha)+\Phi^{-1}(\beta_1))^2}{\theta_R^2}$. A value of 2.5% for α implies that $\Phi^{-1}(\alpha)$ is 1.96 and a value of 80% for $1-\beta_1$ implies that $\Phi^{-1}(\beta_1)$ is 0.8416. For the first 29 observations the model between *W* and *Z* gives an MSE of 440.217. The reference improvement of interest θ_R is 10.0. Substituting these values in Eq. (4) gives $n_1 = 35$. As n_1 is larger than n_2 an additional 6 (35–29) observations are taken.

Fig. 4 gives a scatter plot of the fitted heart rate at 5 min versus the index of the patient under the model incorporated.

Fig. 4 shows that all of the 35 patients have a fitted heart rate in excess of 80 bpm at 5 min.

An ANOVA table similar to Table 2 was obtained for the Gaussian error linear model between diff and bhr fitted to the 35 observation . A normal probability plot similar to Fig. 2 and a residual versus fitted value plot similar to Fig. 3 were also plotted for the case of 35 observations. These show that the assumptions of the Gaussian error linear model are well satisfied. This table and figures are not given in the paper due to space limitations.

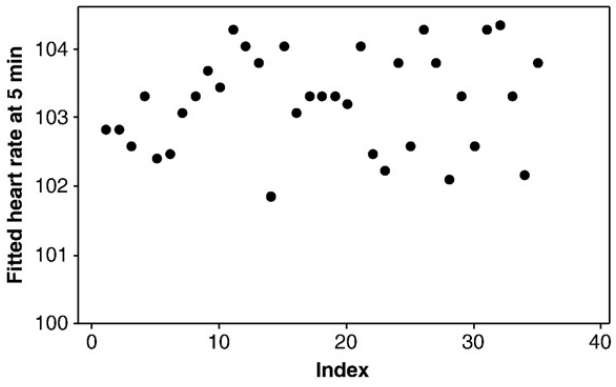

Fig. 4. Scatter plot of fitted heart rate at 5 min Vs. Index for 35 observations.

Based on the total sample of 35 observations, the correlation between diff and bhr was −0.464 (p-value = 0.0019) and the sample mean of diff ($\overline{w}$) was 32.5148 and the mean square error of the first model corresponding to 29 observations was 440.217.

3.3.2. Hypothesis testing

Testing Hypothesis H_{01}:

$$t_0 = \frac{32.5148}{\sqrt{440.217/35}} = 9.16815 (p-value < 0.001). \quad (6)$$

As the p-value is less than 0.001 there is strong evidence to reject H_{01} and conclude that 0.6 mg of Atropine successfully increases the heart rate. At the mean pulse rate the estimate of the mean heart rate at 5 min is 102.51 bpm. (32.51 + 70).

Testing Hypothesis 2:

$$z' = \left(\frac{1}{2} log\left(\frac{1-0.464}{1+0.464}\right) + \frac{0.464}{2\times 34}\right)\sqrt{(35-3)} = -2.804 (p-value = 0.0025). \quad (7)$$

As the p-value is very much less than 0.025 this hypothesis is also rejected. The estimated correlation between diff and bhr is −0.464. This indicates that the difference of heart rate at 5 min and baseline reduces with baseline heart rate.

4. Discussion

4.1. Main conclusions

The methods of Wittes and Brittain [1] and Coffey and Muller [10] were used for utilising an internal pilot study for testing hypotheses pertaining to a Gaussian error linear model for a clinical toxicology study. The method of Stein [27] was used in testing the first hypothesis. This adjusts the type I error rate of the first test for selection between models at the interim stage.

At the interim stage it was found that a model which is very suitable for relating the change in heart rate with basal

heart rate is the Gaussian error linear model between diff and bhr. The MSE of this model was used to calculate the final sample size.

The first hypothesis test indicates that 0.6 mg of Atropine significantly increases the mean heart rate at 5 min after treatment. At the mean pulse rate the estimated mean heart rate at 5 min was 102.51 bpm and the mean increase in heart rate at 5 min and baseline was 32.51. The mean baseline heart rate in this population was considered to be 70 bpm in these calculations. This study suggests that a single 0.6 mg dose of IV atropine is sufficient to treat yellow oleander-induced bradycardia. The second hypothesis test shows that the difference of heart rate at 5 min and baseline reduces with baseline heart rate. This study found that the baseline heart rate is an important variable determining the change in heart rate between, 5 min after injection of Atropine and baseline. This is why the model includes a measure of the baseline heart rate (Z). By including this variable (Z) in the model we adjust for baseline differences. Simulation studies based on this problem show that the type I error rate and power are well maintained within acceptable limits.

4.2. Wider applicability of methods developed

While these adaptations are instrumental in achieving the required power, these do not materially affect the significance level.

In practice in many studies there is a lack of knowledge of nuisance parameters and distributional and modelling assumptions. These problems could be overcome by using an internal pilot study to conduct a review for gaining knowledge about these unknown parameters and assumptions. It has been found [1,2,4,10] that in large scale clinical trials while these adaptations are instrumental in achieving the required power, these do not materially affect the significance level, but the performance of these methods had yet to be established for small scale clinical studies.

In this paper the methods explained have been applied to a small uncontrolled observational study and illustrated that the methods examined are valid for small clinical studies too. While experimental methods (randomised controlled trials) are the "gold standard" for evaluation, observational studies have provided and will continue to make unique and important contributions to the totality of evidence upon which to support a judgment of proof beyond a reasonable doubt in the evaluation of interventions [29]. All the methods discussed here can well be applied to small scale clinical trials where the patients are randomised to two treatment groups. The only issue which should be considered for clinical trials is that of blinding and whether unblinding of treatment allocation at design review could be allowed. For large scale trials Gould [2] and Gould and Shih [4] recommend the use of a blinded method of estimating variance at design review which is based on the EM algorithm. Subsequently simulation studies done by Sooriyarachchi [30] revealed that this estimate is very biased for small samples. These results have been supported by the findings of Marschner [23]. Marschner has developed a modified version of the EM algorithm for estimating the variance without unblinding and this seems promising though more simulation studies are required before it can be recommended. Sooriyarachchi [30] also studied the performance of a Bayesian version of the EM algorithm, namely the MCEM algorithm for estimating the variance at design review without unblinding treatment allocation. Sooriyarachchi [30] found that this method gives promising results for small samples when there is good previous evidence to base a guess of variance upon. This though is not the case when previous evidence is poor. When there is a lot of uncertainty about nuisance parameters considering unblinded estimation of these parameters at design review and overcoming any procedural issues is supported by certain authors [31].

5. Contributions

Dr. Michael Eddleston set up the observational cohort with Dr. Hasantha Ranganath who collected the data for this study. Professor Roshini Sooriyarachchi designed the statistical approach to the study design, conducted the review and simulation study and wrote up this paper. Ms. Rasika Jayatillake assisted in the analysis and prepared the figures and tables.

Acknowledgements

The authors are grateful to the South Asian Clinical Toxicology Research Collaboration (SACTRC) for permission to publish these data and in particular would like to thank Professor Andrew Dawson for this collaboration.

References

[1] Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Statistics in Medicine 1990;9:65–72.
[2] Gould AL. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. Statistics in Medicine 1992;11:55–66.
[3] Gould AL. Planning and revising the sample size for a trial. Statistics in Medicine 1995;14:1039–51.
[4] Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics, Theory and Methods 1992;21(10):2833–53.
[5] Zucker DM, Wittes JT, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedures. Statistics in Medicine 2001;20:2625–43.
[6] Shih WJ. Sample size re-estimation-journey for decade. Statistics in Medicine 2001;20:515–8.
[7] Gould AL. Sample size re-estimation: recent developments and practical considerations. Statistics in Medicine 2001;20:2625–43.
[8] Whitehead J, Whitehead A, Todd S, Bolland KM, Sooriyarachchi MR. Mid-trial design reviews for sequential clinical trials. Statistics in Medicine 2001;20:165–76.
[9] Bolland KM, Sooriyarachchi MR, Whitehead J. Sample size review in head injury trial with ordered categorical responses. Statistics in Medicine 1998;17:2835–47.
[10] Coffey CS, Muller KE. Exact test size and power of a Gaussian error linear model for an internal pilot study. Statistics in Medicine 1999;18:1199–214.
[11] Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. Biometrical Journal 2006;48(4):537–55.
[12] Institute of Medicine. Small Clinical Trials: Issues and Challenges. The National Academic Press; 2001.
[13] Eddleston M, Ariaratnam CA, Meyer PW, Perera G, Kularatne SAM, Attapatu M, et al. Epidemic of self-poisoning with seeds of the yellow oleander tree (Thevetia peruviana) in northern Sri Lanka. Trop Med Int Health 1999;4:266–73.
[14] Roberts DM, Eddleston M. Yellow oleander poisoning. Critical Care Update 2004; 2004. p. 189–200. New Delhi, Jaypee.
[15] Eddleston M, Ariaratnam CA, Sjostrom L, Meyer WP, Perera G, Kularatne SAM, et al. Acute yellow oleander (Thevetia peruviana) poisoning — cardiac arrhythmias, electrolyte disturbances, and serum cardiac glycoside levels on presentation to hospital. Heart 2000;83:301–6.

[16] De Silva HA, Fonseka MMD, Pathmeswaran A, Alahokone DGS, Ratnatilake GA, Gunatilake SB, et al. Multiple-dose activated charcoal for treatment of yellow oleander poisoning: a single-blind, randomised, placebo-controlled trial. Lancet 2003;36:1935–8.
[17] http://curriculum.toxicology.wikispaces.net/Course + development. Developed in July, 2006. Accessed in July 2010.
[18] Rajapakse S. Management of yellow oleander poisoning. Clinical Toxicology 2009;47(3):206–12.
[19] Ali-Melkkia T, Kanto J, Lisalo E. Pharmacokinetics and related pharmacodynamics of anticholinergic drugs. Acta Anaesthesiol Scandinavia 1993;37:633–42.
[20] http://www.google.lk/url?q=http://curriculum.toxicology.wikispaces.net/file/view/cardiac%2Bglycoside%2Bmechanisms2.ppt&sa=U&ei. Developed in 2007. Accessed in July 2010.
[21] Machin D, Campbell MJ, Fayers PM, Pinol APY. Sample Size Tables For Clinical Studies2nd edn. . Oxford: Blackwell Science; 1997.
[22] Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1983.
[23] Marschner I. Miscellanea: on stochastic versions of the EM algorithm. Biometrika 2001;88(1):281–6.
[24] Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1958.
[25] Chow S, Chang M. Adaptive Design Methods in Clinical Trials. USA: Chapman and Hall/CRC Biostatistics Series; 2006.
[26] Draper NR, Smith H. Applied Regression Analysis. 3rd edn. Singapore: Wiley; 2003.
[27] Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. Annals of Mathematical Statistics 1945;16:243–58.
[28] Kanji GK. 100 Statistical tests. SAGE publication; 2000.
[29] Black N. Why we need observational studies to evaluate the effectiveness of health care. British Medical Journal(BMJ) 1996;312:1215–8.
[30] M.R. Sooriyarachchi, 'The use of interim inspections for decision making in clinical trials' unpublished Ph.D. thesis, University of Reading, 1994.
[31] http://www.biopharmanet.com accessed on 28.04.10.