A three-stage clinical trial design for rare disorders

Visa E. A. Honkanen¹,²,³, Andrew F. Siegel⁴, John Paul Szalai⁵, Vance Berger⁶, Brian M. Feldman¹,²,⁷,*,† and Jeffrey N. Siegel⁸

¹Division of Rheumatology, The Hospital for Sick Children, Toronto, Canada
²University of Toronto, Toronto, Ontario, Canada
³Children's Hospital, University of Helsinki, Helsinki, Finland
⁴Departments of Management Science, Finance, Statistics and Molecular Biotechnology, University of Washington at Seattle and the Institute for Systems Biology, Seattle, U.S.A.
⁵Department of Research Design and Biostatistics, Sunnybrook Health Science Center, Toronto, Canada
⁶Food and Drug Administration, Center for Biologics Evaluation and Research, Office of Establishment Licensing and Product Surveillance, Division of Biostatistics and Epidemiology, Rockville, U.S.A.
⁷Department of Paediatrics and Community Medicine, University of Toronto, Toronto, Canada
⁸Food and Drug Administration, Center for Biologics Evaluation and Research, Office of Therapeutics Research and Review, Division of Clinical Trial Design and Analysis, Rockville, U.S.A.

SUMMARY

Many clinical trials of uncommon diseases are underpowered because of the difficulty of recruiting adequate numbers of subjects. We propose a clinical trial design with improved statistical power compared to the traditional randomized trial for use in clinical trials of rare diseases. The three-stage clinical trial design consists of an initial randomized placebo-controlled stage, a randomized withdrawal stage for subjects who responded, and a third randomized stage for placebo non-responders who subsequently respond to treatment. Test level and power were assessed by computer-intensive exact calculations. The three-stage clinical trial design was found to be consistently superior to the traditional randomized trial design in all cases examined, with sample sizes typically reduced by 20 per cent to 30 per cent while maintaining comparable power. When a treatment clearly superior to placebo was considered, our design reached a power of 75 per cent with a sample of 21 patients compared with the 52 needed to attain this power when only a randomized controlled trial was used. In situations where patient numbers are limited, a three-stage clinical trial design may be a more powerful design than the traditional randomized trial for detecting clinical benefits. Copyright © 2001 John Wiley & Sons, Ltd.

INTRODUCTION

To assure that a clinical trial will produce interpretable results, the combination of sample size and design must provide adequate statistical power to detect the therapeutic benefit of
a new therapy. When studying rare disorders, the number of patients available to study is usually small, making it difficult to enrol enough patients to power a randomized clinical trial adequately. Another strategy for increasing statistical power is to design the trial to collect more information from each patient, for example, with cross-over designs. Cross-over designs, however, rely on every patient taking part in a placebo stage equal in length to the treatment stage and on a total washout of the treatment effect, without which the design may yield invalid results [1, 2]. In addition, when expectations for the drug are high and equipoise is lacking [3], the placebo stage may pose a recruitment problem [4]. These problems become especially evident in studies of rare diseases in paediatric populations, such as juvenile rheumatoid arthritis or juvenile dermatomyositis [5, 6], when the treatment may have been tested in adults with promising results, thus raising expectations for its efficacy in children.

Another alternative to the simple randomized trial is the randomized withdrawal design [7]. A randomized withdrawal trial is an example of an enrichment design in which subjects previously demonstrated to respond to treatment are randomly withdrawn to placebo or maintained on therapy. Recently, the withdrawal design has been used successfully in studies of drugs with long or uncertain wash-out times in chronic disorders [8, 9] and was provided as evidence of efficacy for the approval of etanercept for juvenile rheumatoid arthritis [10]. We asked whether by combining a simple randomized trial with a randomized withdrawal trial in the same group of subjects, thereby getting more information from each one, we could study fewer patients than in a traditional randomized control trial design. We used this approach to develop a three-stage clinical trial design, consisting of an initial randomized placebo-controlled stage, a randomized withdrawal stage for subjects who responded to drug, and a third randomized stage for initial non-responders to placebo who subsequently responded to drug treatment given open-label. We then assessed the power of this design using computationally-intense exact calculations to ascertain whether it worked, whether it offered practical advantages over more conventional designs, and whether studying the same subjects twice led to false positive results.

METHODS

Three-stage clinical trial design

The design was based on classifying the treated subjects as responders or non-responders, a method now recommended in analysing the efficacy of treatment in a number of chronic rheumatic conditions. The American College of Rheumatology has published standardized criteria for a significant clinical response to a drug therapy in adults [11], and proposed a similar type of criteria for children. To assess responses in the randomized withdrawal stage of the trial, criteria for flare of disease activity are used. Examples of criteria for flare include a published trial of randomized withdrawal of hydroxychloroquine in systemic lupus erythematosus [12] or a trial utilized to determine the efficacy of etanercept in juvenile rheumatoid arthritis [10]. In the discussion which follows, a responder in the randomized withdrawal stage of the trial is defined as a subject who does not flare.

Our design is illustrated in Figure 1. In the first stage, patients are enrolled into a standard two-group parallel-arm randomized clinical trial. Each patient is randomized to receive either
Figure 1. Design of the three-stage design clinical trial. Stage I of the three-stage design consists of an ordinary randomized placebo-controlled trial, which yields the first two-by-two contingency table with $p_1$ as its one-sided chi-squared $p$-value. In the second stage, the patients who responded to treatment in stage I are randomly assigned to treatment or placebo, resulting in the stage II two-by-two table with $p_2$ as its one-sided chi-squared $p$-value. In the third stage, the patients who did not respond to placebo in stage I are placed on active treatment, and responders are randomly assigned to treatment or placebo, resulting in the stage III two-by-two table with $p_3$ as its one-sided chi-squared $p$-value. Variable names in each circle indicate the number of subjects, for example, $n_{CNTR}$ is the number of control (placebo) non-responders subsequently assigned to treatment who then responded. Probabilities of responding are indicated along branches, for example, $p_{TC}$ is the probability of responding by a subject initially assigned to the treatment group but subsequently reassigned to the control (placebo) group. We assume that each subject responds independently at the appropriate probability at each stage so that, for example, $nTR$ follows a binomial distribution with parameters $nT$ and $pT$. Assignments are made in a balanced way so that, for example, $nT = \text{int}(n/2)$ and $nC = n - nT$. The one-sided chi-squared $p$-value, for which smaller values correspond to more effective treatment, may be computed from the conventional two-sided chi-squared $p$-value as follows. If subjects assigned to treatment respond at a higher rate than those assigned to control, then set $p_{\text{one-sided}} = p_{\text{two-sided}}/2$. If they respond at a lower rate, then set $p_{\text{one-sided}} = 1 - p_{\text{two-sided}}/2$. If they respond at an equal rate, then set $p_{\text{one-sided}} = 1/2$. In addition, if computation of the chi-squared statistic would involve division by zero, then set $p_{\text{one-sided}} = 1/2$. The drug studied (active therapy) or a placebo. Subjects may either respond or not; the results produce the stage I two-by-two contingency table whose one-sided chi-squared $p$-value is denoted $p_1$. At this point, the subjects who did not respond to active treatment end the study.
(in a cross-over trial they would start on placebo after a wash-out period.) Similarly, subjects who respond favourably to placebo also end the study (in a cross-over trial they would start on active treatment after a wash-out period).

In the second stage, the patients who did \textit{did} respond to treatment in stage I are randomized either to stay on active treatment or withdraw and be placed on placebo. In these patients the disease either recurs or stays in remission. The original treatment arm ends with the stage II two-by-two contingency table whose one-sided chi-squared \( p \)-value is denoted \( p_2 \).

In the third stage, the patients who did \textit{not} respond to placebo in stage I are placed on active treatment: those who do not respond exit the study, and responders are randomly assigned either to stay on active treatment or return to placebo. In these patients, the disease either recurs or stays in remission. The original placebo arm ends with the stage III two-by-two contingency table whose one-sided chi-squared \( p \)-value is denoted \( p_3 \).

The \( p \)-values from the three stages \((p_1, p_2 \text{ and } p_3)\) are asymptotically independent and asymptotically distributed according to a uniform distribution on \((0,1)\) under the hypothesis of no differences between treatment and placebo in any of the three stages, so that \( p_T = p_C \), \( p_{TT} = p_{TC} \), and \( p_{CTT} = p_{CTC} \), using the notation in Figure 1. This distribution follows from the fact that independent random assignments to treatment and placebo are made at each stage and from asymptotic properties of chi-squared tests. It is worth noting that these results remain true even when we re-use the subjects because each two-by-two table is independently randomized. These three \( p \)-values are then combined using Fisher’s method, which rejects the null hypothesis if the product \( p_1 p_2 p_3 \) is less than 0.001844. (This may be derived from the fact that \(-2 \ln(p_1 p_2 p_3)\) follows, asymptotically, a chi-squared distribution with 6 degrees of freedom.) If the two-by-two table at either stage II or stage III would involve fewer than four people, it is not used, and the result of the three-stage design would be computed either using Fisher’s method with the two available \( p \)-values (rejecting the null hypothesis if the appropriate product \( p_1 p_2 \) or \( p_1 p_3 \) is less than 0.008705; this may be derived from the fact that \(-2 \ln(p_1 p_2)\) follows, asymptotically, a chi-squared distribution with 4 degrees of freedom) or using just \( p_1 \) if it is the only \( p \)-value available.

Fisher’s method was chosen over two alternative methods for combining \( p \)-values that were considered. The first alternative, the minimum \( p \)-value (using critical value \((1 − 0.95)^k\) when \( p \)-values from \( k \) stages are combined), was found to have uniformly lower power than Fisher’s method in all scenarios considered. The second alternative, a weighted Fisher’s method (using \(-2 \ln(p_1^n p_2^m p_3^k) = -2n_1 \ln(p_1) - 2n_2 \ln(p_2) - 2n_3 \ln(p_3)\) as the test statistic, where \( n_i \) is the number of subjects enrolled in stage \( i \)), gave similar results to Fisher’s method in many cases, but was considerably worse than Fisher’s method in the mixture model (for example, having 67 per cent power as compared to 84 per cent for Fisher’s method in the case of initial sample size 25). We therefore chose the unweighted Fisher’s method as a simple, effective and well-established solution.

Please note that we are not claiming that the statistical data are independent from one stage of the trial to another, and such a claim is not necessary for the success of the three-stage design. Instead, we assert the weaker claim that the \( p \)-values are asymptotically independent across the three stages (even though the response rates, for example, might not be independent). For example, if the stage I treatment group happens randomly to show a high response rate, we may well see higher response rates in stage II (consistent with statistical dependence between the results of the two stages); none the less, under the hypothesis of no treatment effect, the \textit{full and complete re-randomization} of all subjects who become part of stage II
will cause the $p$-value for stage II to be asymptotically uniformly distributed regardless of the results of stage I. Since the stage II $p$-value has the same asymptotic conditional probability distribution (uniform $[0,1]$) regardless of the result of stage I, it follows that the stage II $p$-value is asymptotically statistically independent of the stage I $p$-value. This is all that is needed for the asymptotic validity of Fisher’s method for combining independent $p$-values from the three stages.

Note that the analysis proposed here assumes random marginals in the tables, that is, the total number of subjects responding at a given stage is taken to be random. While this is a well-established approach, one alternative is to assume that the total number responding is fixed, and the randomness arises only from the randomized assignments [13]. This could be accomplished here by substituting one-sided $p$-values from Fisher’s exact test in place of $p$-values from the chi-squared test while leaving all other aspects of the design unchanged.

**Evaluating the design**

We developed software to evaluate the performance of the full three-stage design and of the traditional randomized clinical trial design (which uses only the first stage) by computationally-intensive enumeration of all possible outcomes that could occur, given the initial sample size $n$ and values for the seven probabilities of responding ($p_T, p_C, p_{TT}, p_{TC}, p_{CT}, p_{CTC}$) as shown in Figure 1 (for example, with $n=40$ there are 5473556 possible outcomes). For each possible observable outcome (defined as the seven numbers $n_{TR}, n_{CR}, n_{TRTR}, n_{TRCR}, n_{CNTR}, n_{CNTRTR}$ and $n_{CNTRCR}$ each specifying a number of subjects responding) we computed the stage I (randomized clinical trial) $p$-value $p_1$ and the combined result for all three stages to see if the null hypothesis was rejected at the 5 per cent level. We also, for each possible outcome, computed its probability of occurrence as the product of the seven binomial probabilities, one corresponding to each probability of responding. By summing these probabilities of occurrence for the observed outcomes for which the null hypothesis was rejected, we were able to compute the exact probability of rejecting the null hypothesis for each design.

Here is the algorithm we used to enumerate the sample space and compute the exact power of a test:

1. Set the initial number assigned to treatment: $n_T = \text{Int}(n/2)$, where Int is the integer part function.
2. Set the initial number assigned to control: $n_C = n - n_T$.
3. Set the number of stage I treatment responders using a control loop: FOR $n_{TR} = 0$ TO $n_T$.
4. Set the number assigned to treatment in stage II: $n_{TRT} = \text{Int}(n_{TR}/2)$.
5. Set the number assigned to control in stage II: $n_{TRC} = n_{TR} - n_{TRT}$.
6. Set the number of stage I control responders using a control loop: FOR $n_{CR} = 0$ TO $n_C$.
7. Assign all control non-responders to treatment in preparation for stage III: $n_{CNT} = n_C - n_{CR}$.
8. Set the number of treatment responders in stage II using a control loop: FOR $n_{TRTR} = 0$ TO $n_{TRT}$.
9. Set the number of control responders in stage II using a control loop: FOR $n_{TRCR} = 0$ TO $n_{TRC}$.
10. Set the number of subjects who will participate in stage III (these are stage I control non-responders subsequently responding to treatment) using a control loop: FOR \( nCNTR = 0 \) TO \( nCNT \).

11. Set the number assigned to treatment in stage III: \( nCNTRT = \text{Int}(nCNTR/2) \).

12. Set the number assigned to control in stage III: \( nCNTRC = nCNTR - nCNTRT \).

13. Set the number of treatment responders in stage III using a control loop: FOR \( nCNTRT = 0 \) TO \( nCNTRT \).

14. Set the number of control responders in stage III using a control loop: FOR \( nCNTRCR = 0 \) TO \( nCNTRC \).

15. Within this innermost loop, we see every possible configuration (for example, the stage I table counts are \( nTr, nT - nTR, nCR, nC - nCR \)). We can therefore compute all appropriate \( p \)-values and accumulate the probability if the null hypothesis is rejected. The probability of the current configuration (reflecting all three stages) may be computed as the following product of binomials and conditional binomials (although we optimized for efficiency by precomputing binomial probabilities and moving multiplications outside control loops wherever possible):

\[
\begin{align*}
{nT \choose nTR} p^{nT}(1 - pT)^{nT-nTR} {nC \choose nCR} p^{nC}(1 - pC)^{nC-nCR} \\
\times {nTRT \choose nTRTR} p^{nTRT}(1 - pTR)^{nTRT-nTRTR} \\
{nTRC \choose nTRCR} p^{nTRC}(1 - pTC)^{nTRC-nTRCR} {nCNT \choose nCNTR} p^{nCNT}(1 - pCT)^{nCNT-nCNTR} \\
\times {nCNTRT \choose nCNTRTR} p^{nCNTRT}(1 - pCTR)^{nCNTRT-nCNTRTR} \\
{nCNTRC \choose nCNTRCR} p^{nCNTRC}(1 - pCTR)^{nCNTRC-nCNTRCR}
\end{align*}
\]

16. All control loops end here: NEXT \( nCNTRCR \), NEXT \( nCNTRTR \), NEXT \( nCNTR \), NEXT \( nTRCR \), NEXT \( nTRTR \), NEXT \( nCR \), NEXT \( nTR \).

RESULTS

In a standard randomized clinical trial, efficacy of a therapeutic agent is inferred from the initial response of patients to treatment compared to control. In the three-stage design (Figure 1), the trial begins with an initial exposure to drug or control (stage I) identical to a standard randomized clinical trial. Then responders to drug are randomized to remain on drug or be withdrawn to placebo (stage II). In addition, subjects who did not respond to placebo in the initial stage are given drug open-label. Those who respond are also randomized to remain on drug or be withdrawn to placebo (stage III). Thus, in contrast to the standard randomized controlled trial which infers efficacy of a new therapy only based on the initial exposure
to drug or placebo, the three-stage design also uses information from the two withdrawal stages (stages II and III) to reach inferences about the efficacy of the therapy. We studied various scenarios in which this new design might be used. Response rates for the active treatments were based on the expectation that the response rates of active therapy would be 0.4 and 0.5, that is, 40 per cent or 50 per cent of the patients would be responders. These are approximately the response rates achieved with some of the new biologic therapies being studied in rheumatic diseases. Since our design is intended for studies of chronic, severe diseases where the response rate to placebo is usually low, we set the response rate to placebo at 20 per cent when the active treatment was superior to placebo.

Type I error measures the chance that a clinical trial will show that a therapy is efficacious even though there is actually no difference between treatment and control. Figure 2 shows the exact probability of type I error when the nominal test level is 5 per cent. Results are shown for each method (stage I RCT and three-stage design) for a uniform response rate of 20 per cent at each sample size (20, 25, 30, 35, 40, 45, 50, 55, 60 and 65) together with results for an uniform response rate of 40 per cent at each sample size. Some deviation from the nominal level is to be expected due to finite sample size. The results indicate that the three-stage design is not associated with a statistically higher probability of type I error than the standard stage I randomized trial.

The statistical power of a clinical trial is defined as the ability of a trial to detect a treatment effect if the treatment is effective, given a specified sample size and certain assumptions about the effects of therapy. Figure 3 shows the statistical power of the three-stage design and the randomized clinical trial design when the response rate to treatment is 40 per cent and the response rate to placebo is 20 per cent, for sample sizes between 20 and 65 (20, 25, 30, 35, 40, 45, 50, 55, 60 and 65). The three-stage design was consistently more powerful than a standard randomized clinical trial for all sample sizes studied. Under these assumptions, 50 per cent power is achieved with a sample size of 49 subjects with the three-stage design.
Figure 3. Statistical power of the three-stage design assuming a response rate of 40 per cent. Exact statistical power of the three-stage design and the randomized clinical trial design when the response rate to treatment is 40 per cent and the response rate to placebo is 20 per cent, for sample sizes 20, 25, 30, 35, 40, 45, 50, 55, 60 and 65.

Figure 4. Statistical power of the three-stage design assuming a response rate of 50 per cent. Exact statistical power with response rate to treatment set at 50 per cent while the response rate to placebo is 20 per cent.

compared to 54 subjects using a standard randomized controlled trial. Figure 4 shows how these results change when the treatment response rate is increased to 50 per cent. The three-stage design was again consistently more powerful than the standard randomized clinical trial at all sample sizes. Under these assumptions, 75 per cent power is achieved with a sample size of 42 subjects with the three-stage design compared to 52 subjects using a standard, randomized controlled trial.

Since the subjects in the randomized withdrawal stages of the trial (stages II and III) are previously selected for having responded to drug (in stage I), the proportion of the subjects who maintain a response to drug may be higher than the proportion who respond to drug in the initial stage of the trial. This inequality in expected response rates can be modelled using
Figure 5 shows the statistical power for a non-uniform response rate model in which response rates are allowed to differ in each stage. For this purpose, we assumed a mixture model for the population, in which 30 per cent of subjects always respond to treatment and never respond to placebo, while the remaining 70 per cent respond randomly and independently at a constant rate regardless of assignment to treatment or placebo. Stage I response rates are set at 50 per cent for treatment and 20 per cent for placebo. Selection effects increase the probability of responding to treatment and decrease the probability of responding to placebo in stages II and III, yielding increased power.

In a clinical trial conducted using the three-stage design, the results of treatment with study drug (stage I) and randomized withdrawal (stages II and III) give rise to three separate assessments of efficacy, in contrast to the single measure of efficacy used to calculate overall statistical significance in a conventional randomized clinical trial. The individual measures of
statistical significance are then combined using Fisher’s method to derive a single overall $p$-value. The three-stage design was found to be uniformly more powerful than the single-stage randomized clinical trial in all cases studied, especially in trials with low patient numbers and highly effective treatment. Despite the increase in power, the three-stage design is not associated with a higher probability of type I error than the conventional parallel-arm randomized clinical trial design. As an example of the relative power of the three-stage design, when the response rates were set uniformly at 50 per cent for active treatment and 20 per cent for placebo, statistical power of 75 per cent is attained at 42 patients by using the three-stage design. In a single-stage design we would need 52 patients to achieve the same statistical power (Figure 4). If the initial response rates are 50 per cent for active treatment and 20 per cent for placebo, but the population is a mixture of consistent and random responders, a power of 75 per cent is achieved by the three-stage design at 21 patients, less than half of the 52 necessary with a conventional randomized clinical trial (Figure 5).

Several caveats should be considered in the interpretation of the second and third stages of the trial that involve drug withdrawal. First, before subjects are enrolled in the withdrawal stage, they must first have responded to study drug, so that the frequency of flare in the withdrawal study must be understood to pertain to the responding subpopulation and not to the entire patient population. However, one of the advantages of the three-stage design over some withdrawal designs is that the population of patients randomized in the withdrawal stages consists of all subjects who initially responded. Second, in withdrawal studies, there is the possibility that drug withdrawal itself may induce a flare of disease that is greater than the patient had at baseline. The three-stage design should be avoided in the assessment of drugs known to cause a rebound flare of disease when they are withdrawn. To conclude that a positive result in a three-stage clinical trial implies efficacy of the study drug, it would be important to observe trends indicating an effect of drug in the induction stage of the trial (stage I) as well as in the withdrawal stages (stages II and III).

Similar to other clinical trials, a clinical trial using the three-stage design could incorporate an interim analysis with early stopping rules if efficacy is seen after stage I has been completed. If the $p$-value for stage I ($p_1$) is small enough, there would be no reason to complete stages II and III for purposes of hypothesis testing. In addition, a penalty would not be necessary because one cannot get significance after stage I and lose it later, by virtue of the fact that $p_1 < p_2 < p_3 < p_1$. However, investigators may wish to complete the trial regardless, since the withdrawal stages of the trial provide additional information about whether continued treatment with the drug is necessary to maintain the clinical effects. An interim analysis for futility could also be incorporated after stage I so that if the treatment is found to lack efficacy subjects can be spared moving into stages II and III unnecessarily.

The three-stage design offers several features which may address the concerns of subjects and their physicians about the unnecessary exposure to placebo. In our design, all the patients entered in the study have a chance to derive a therapeutic benefit from their participation in the trial in that all subjects assigned to the placebo arm who do not improve are subsequently offered active treatment in the later stages of the trial. In addition, the three-stage design avoids treating patients unnecessarily in that the subjects who respond to placebo in the first stage do not undergo active treatment. Similarly, those who do not respond to active treatment in the first stage do not have to go on placebo. In addition, only half of the patients who respond to active treatment in the first stage have to change from treatment to placebo. As an example, if the response rate to treatment is 50 per cent and to
THREE-STAGE TRIAL DESIGN FOR RARE DISORDERS 3019

placebo 20 per cent, and the number of patients is 50, then instead of 25 (cross-over design), only 11–12 patients on average have to start on placebo after the first stage, and 5 patients are saved from active treatment. (This average of 11–12 may be found by adding the expected number of patients enrolled in placebo in stages II and III as follows: \( \frac{n}{2}(p_T) + \frac{n}{2}(1-p_C)(p_CT) \), using \( n = 50 \), \( p_T = p_CT = 0.5 \) and \( p_C = 0.2 \).) These features in our design may make it more acceptable to informed patients without compromising the statistical rigour of the trial.

The challenge in conducting clinical trials of rare disorders is to enrol enough patients to reach acceptable statistical power. Sometimes even a multi-centre collaboration does not provide a sufficient number. Too often, researchers try to solve this problem by overestimating the number of patients to be enrolled or simply go ahead, hoping that despite low statistical power they will attain a significant result. Thus, a study with a small number of patients runs the risk of failing to detect even a highly effective treatment [14]. Conducting a trial with low power or without an adequate power analysis could be considered unethical. A cross-over design gives the best possible power when all the conditions are met [15]. Often this is not the case, since to obtain unbiased results, the intervention must have a completely reversible effect [2]. With many interventions, however, we do not know whether this is true. In addition, when \( a \) priori expectations of the intervention are strongly positive, recruitment of patients into a full-scale cross-over trial becomes difficult [2]. The sparseness of randomized clinical trials in paediatric rheumatological disorders, such as juvenile dermatomyositis, illustrates these problems [16].

Our design does not solve all these problems but can provide substantial advantages in certain situations. First, the patients changed from placebo to treatment are not included in the same analysis with those going from treatment to placebo. Hence, the potential carry-over effect from treatment does not compromise the results. Second, all the patients entered in the study have a chance to receive the active treatment. This factor, together with saving some patients from an unnecessary treatment or placebo stage, may make the design more attractive to subjects, especially in a situation where preliminary data suggest that the treatment is effective. This is often the case in paediatric medicine. When the treatment effect was clearly superior to placebo, it is possible to reach an acceptable power with 21 patients instead of the 52 who would have been needed with the classic randomized controlled trial design. Differences of this magnitude are important when dealing with rare disorders. Our design is more complicated to work through than a simple parallel trial, but this complexity must be weighed against the power advantage it offers.

It should be noted that the three-stage design is not optimal for all clinical situations and does not replace the traditional randomized trial design. First, it is applicable only to chronic conditions where both response to therapy and flare upon withdrawal of therapy can be assessed. So, for example, treatments with clinical effects that do not wash out within a reasonable time frame would probably not be good candidates for this type of design. In designing a three-stage clinical trial, care should be taken to allow a duration of the withdrawal phase long enough so that the drug would be completely washed out and the clinical effects of therapy reversed, otherwise the flare rates in the withdrawal phase may not approach the response rate in stage I. However, unlike the cross-over design, there is no expectation that subjects return to their pre-treatment clinical status during stages II and III, since a flare can be defined in a way that represents just a partial return toward baseline. Second, some placebo subjects may barely meet criteria for being a responder and would consequently forgo active
treatment even though they may have benefited from it. To minimize the number of spurious placebo responses, care should be taken in selecting clinically relevant response criteria that minimize the frequency of placebo responders. Third, since fewer patients may be available in the initial stage of the trial, the ability to precisely determine initial response rates may be less than with a traditional randomized trial design. Fourth, the three-stage design may be less suited for controlled assessment of safety in that the initial stage of randomized assignment to active treatment or control is limited in patient number and in duration. None the less, because of its advantages in power, the three-stage design may be particularly helpful in studies of rare diseases where subject number is limited. Another situation where this study design may offer advantages is in the efficacy of a therapeutic agent in a particular patient subpopulation when efficacy in the general patient population has already been established. Finally, the three-stage design may be helpful in early stages of drug testing where small cohorts of patients are tested but where decisions must be made about the choice of dosing, about choosing between alternative candidate compounds or about whether to proceed with drug development.

In recent years, as patients have become more aware of the methods used in clinical trials, randomized studies have been criticized on ethical grounds [17, 18]. This perception may affect both recruitment and compliance. We can address this problem in two ways. First, we can emphasize to patients and the general public the importance of objective studies, even of promising therapies; the double-blind randomized design remains the gold standard for clinical trials. Second, we can try to tailor designs to each situation and test the validity, power and robustness of each one before using the model in a real trial.

Our design can be used in situations in which the number of patients is less than 60, and preliminary data indicate that the treatment is at least moderately effective. In these situations, the same statistical power can be achieved with fewer patients, cutting required sample size by about 20 per cent to 30 per cent in many cases and up to 60 per cent in others.

ACKNOWLEDGEMENTS

Andrew F. Siegel holds the Grant I. Butterbaugh Professorship at the University of Washington. Brian M. Feldman is an Ontario University of Health Career Scientist.

REFERENCES


