Lupus (1999) 8, 612–619
© 1999 Stockton Press. All rights reserved 0961-2033/99 $15.00
http://www.stockton-press.co.uk/lup

PAPER

Sample size considerations for superiority trials in systemic lupus erythematosus (SLE)

Andrew D Moore[1]* and Lawrence Joseph[2,3]
[1]Department of Clinical Immunology and Allergy, Department of Medicine, The Montreal General Hospital, McGill University, Montreal, Quebec, Canada; [2]Division of Clinical Epidemiology, Department of Medicine, The Montreal General Hospital, McGill University, Montreal, Quebec, Canada; and [3]Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada;

For reasons of efficiency and ethics, sample size calculations are an important part of the design of all clinical trials. This paper highlights the statistical issues inherent to the estimation of sample size requirements in superiority trials particular to SLE. Calculations based on statistical power for testing hypotheses have historically been the method of choice for sample size determination in clinical trials. The advantages of using confidence intervals (CI's) rather than *P*-values in reporting results of clinical trials is now well established. Since the design of a trial should match the analysis that will eventually be performed, sample size methods based on ensuring accurate estimation of important parameters via sufficiently narrow CI widths should be preferred to methods based on hypothesis testing. Methods and examples are given for sample size calculations for continuous and dichotomous outcomes from both a power and confidence interval width viewpoint. An understanding of sample size calculations in association with expert statistical consultation will result in better designed clinical trials that accurately estimate clinically relevant differences between treatment outcomes, thereby furthering the treatment of patients with SLE.

Keywords: systemic lupus erythematosus; sample size; clinical trials

Introduction

Sample size calculations are an important part of the design of all clinical trials. Superiority trials attempt to establish the clinical superiority of a new therapeutic agent when compared to either the current standard of care or a placebo. Sample size estimation in such trials is both clinically and ethically relevant. While too few subjects may jeopardize the ability of a study to accurately estimate a clinically important difference between two treatment arms, it is also true that by including too many patients, more than the minimum necessary number may be exposed to risks or harm through randomization. In this era of cost restraint, calculating the minimum required number of patients also provides a scientifically valid way of reducing the costs of clinical research and the length of time to the completion of clinical studies.

*Correspondence: Dr Andrew Moore, The Montreal General Hospital, Division of Clinical Epidemiology, 1650 Cedar Avenue, Room L10-421, Montreal, Quebec, H3G 1A4, Canada
Tel: (+1) 514 934 4641; fax: (+1) 514 934 8293

One of the first steps in the planning of a controlled clinical study is the identification of the primary outcome measure or measures of interest. Of course, these measures should be both clinically relevant to the disease and potentially modifiable by at least one of the interventions under study. In patients with systemic lupus erythematosus (SLE), trial design is influenced by multiple factors. The protean manifestations of SLE means that subsets of patients with organ-specific disease may require different therapeutic modalities, and the number of patients available and willing to enter into controlled trials for treatment of any specific organ disease may be limited. Lupus nephritis is one of the few organ specific manifestations of SLE that has been well studied in randomized, controlled clinical trials.[1–3] Other than objective outcome parameters defining nephritis and renal function in these patients,[4] the precise meaning of 'clinically relevant change' in an individual patient with more generalized disease who may be responding to a novel therapy is difficult to determine. Outcome measures have been standardized in other rheumatological diseases, including ankylosing spondylitis[5] and rheumatoid arthritis.[6] For example, a 'responder index' has been developed for rheumatoid

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

arthritis, which enables cross-study comparisons and simplifies the design of clinical research trials.[7] Although a wide variety of outcome measures exist for measuring disease activity, damage and health related quality of life in patients with SLE, no clear consensus has emerged as to the most appropriate measure or set of measures for use in ongoing clinical trials.[8] A marker for a responder index in SLE has been proposed,[9] but has yet to be validated in clinical studies. Due in part to the heterogeneity of this disease and to the lack of a consensus for clinical outcome measures, the construction of a randomized controlled clinical trial studying the superiority of therapeutic agents or interventions is a challenging endeavour. In the context of current development and investigations of newer treatment modalities in SLE, however, the number of such clinical trials is likely to increase substantially.

Although a complete description of the multitude of methods for sample size calculations is beyond the scope of this paper, several general overview articles and textbooks which have appeared in the biomedical literature[10–12] are good starting points for further reading. This paper highlights the statistical issues inherent to estimation of sample size in superiority trials particular to SLE. Sample size calculations based on traditional testing of the null hypothesis of no group difference, leading to *P*-values, has been the method most often used in SLE clinical trials to date. The advantages of using confidence intervals rather than *P*-values in reporting results of clinical trials is now well established,[13,14] with some leading epidemiology journals deciding not to publish *P*-values at all.[15] One can argue, therefore, that sample size methods based on ensuring sufficiently narrow confidence interval widths should be preferred to methods based on hypothesis testing, since the design of a study should match the analysis that will eventually be performed. This is especially true since very different sample sizes can be suggested by the two different methods, even for the same trial.[16] Nevertheless, for completeness, in this paper we present both approaches, beginning with power calculations for sample sizes in hypothesis testing, followed by estimation of sample sizes based on confidence interval widths. We begin with a brief comparison of the inferences available following the calculation of *P*-values and confidence intervals.

Hypothesis tests versus confidence intervals in the planning of clinical trials

In traditional hypothesis testing, one typically starts by asserting a null hypothesis that the outcomes in two different treatment groups are equivalent, at least on average. Of course, one then hopes that this assertion will be contradicted by the data, as measured by a small *P*-value. While clinicians often misinterpret a *P*-value as providing the probability that the null hypothesis is true after accounting for the information provided by the data, this interpretation is far from correct. In fact, the *P*-value is calculated *assuming that the null hypothesis is true*! Given that the null hypothesis is true, the *P*-value simply provides the probability of obtaining a result as or more extreme than that observed in the trial (i.e. further from what the null hypothesis would predict), if the trial were to be repeated over and over, each time with the null hypothesis in fact being true. The *P*-value is neither directly nor indirectly related to the probability that the null hypothesis is correct, and in fact can be many orders of magnitude different from this quantity.[17] Given the unnatural interpretation of a *P*-value, however, it is no wonder that clinicians often misinterpret it as the quantity they would more naturally desire, the probability of the null hypothesis. In fact, only Bayesian analysis is able to provide this latter quantity.[14]

In superiority trials, the null hypothesis states at the outset that the two treatments in question are on average equal ($H_0 : \mu_0 = \mu_1$, where μ_0 refers to the mean of the outcome variable in the control group and μ_1 refers to the mean of this variable in the treatment group), while the alternative hypothesis states that the proposed treatments are not equal ($H_1 : \mu_0 \neq \mu_1$). The Type I error in a statistical hypothesis test refers to the possibility of rejecting a null hypothesis that two treatment groups are equivalent when, in fact, a difference due to the treatment does not exist. In designing studies, the probability of a Type I error, donated by the Greek letter α, is often set to a value of 5%. While this value is conventional, there is nothing special about $\alpha = 5\%$, and often it may be more appropriate to select smaller values for α (e.g. $\alpha = 1\%$) to reduce the possibility of falsely stating that the treatment is effective when, in fact, it is not. As the number of hypothesis tests performed increases, one increases the probability of falsely rejecting at least one true null hypothesis, so that setting a smaller value for α may be considered in these cases.

A Type II error occurs when one fails to reject the null hypothesis when the new treatment, in fact, does have a different effect than the standard or placebo treatment. The probability of a Type II error is denoted by the Greek letter β, and is often set equal to 10% or 20% in the medical literature. The probability of rejecting the null hypothesis when it is, in fact, false is termed the power of the study, which then occurs with probability $1-\beta$. The power of a study depends

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

both on the sample size and on the true average difference on the measurement scale between the treatment and control groups. All else being equal, as sample size increases, β decreases and the power $(1-\beta)$ increases. In an unbalanced study, there is a decline in power as the number of patients in one group is increased at the expense of the second group so that, for a given total number of patients, the power of a superiority trial is maximized when there are equal numbers of patients in the two study groups. Power may also be increased by including repeated measurements in the same patients, and by using outcome measures that have smaller standard deviations.

The failure of a clinical study to reject a null hypothesis may be due to the use of insensitive outcome measures or to sample sizes that are too small.[18] Felson *et al*[19] reviewed a number of early randomized clinical trials in SLE which compared the use of steroids with immunosuppressive drugs versus steroids alone in the treatment of lupus nephritis and determined that multiple false negative conclusions were reached because of small sample sizes (less than or equal to 50 patients per study). Only by pooling this data were they able to show a statistically significant benefit to combined therapy in the treatment of lupus nephritis. The authors also performed a power analysis which showed that, for a study to prove that an immunosuppressive agent is 50% superior to steroids alone in preventing renal deterioration, 100 high-risk patients would need to be enrolled.

In any study in which the null hypothesis is not rejected, it is always important to calculate a confidence interval for the true treatment difference, in order to draw correct conclusions. Non-rejection of a null hypothesis may be due to there truly being no important difference between the two treatment arms of a trial, or may be due to lack of power, or a combination of these reasons. The *P*-value alone does not allow one to distinguish between these very different conclusions, but a confidence interval does. For example, a confidence interval whose upper and lower limits are both near the null value of a zero treatment difference shows that there is likely to be no important difference between the treatments. A wider confidence interval, however, may not necessarily imply no difference, as potentially important findings are not ruled out. Hence this latter finding is more properly interpreted as inconclusive, rather than negative, despite the 'non-significant' *P*-value. In general, confidence intervals provide more information than *P*-values, since they focus attention on the range of values compatible with the data, on a scale of direct clinical interest. Given a confidence interval, one can assess the clinical meaningfulness of the result, as can be seen in Figure 1. Depending on where the upper and lower confidence interval limits fall in relation to the upper and lower limits of the region of clinical equivalence, different conclusions should be drawn. The region of clinical equivalence, sometimes called the region of clinical indifference, is the region inside of which two treatments would be considered the same for all practical purposes. The point 0, indicating no difference in results between two treatments, is usually included in the region of clinical equivalence, but values above and below 0 are usually also included. How wide this region is depends on each individual clinical situation. For example, if a new treatment is very costly or has important side-effects, a large benefit in the main outcome would be required before this drug becomes an attractive choice compared to current therapy leading to a wide region of clinical equivalence.

Figure 1 summarizes the five different conclusions that can be made after a confidence interval has been calculated:

(1) The CI includes zero, and both upper and lower CI limits, if they were the true values, would not be clinically interesting. Therefore, this variable has been shown to have no important effect.
(2) The CI includes zero, but one or both of the upper or lower CI limits, if they were the true values, would be interesting clinically. Therefore, the results of this variable in this study is inconclusive, and further evidence needs to be collected.
(3) The CI does not include zero, and all values inside the upper and lower CI limits, if they were the

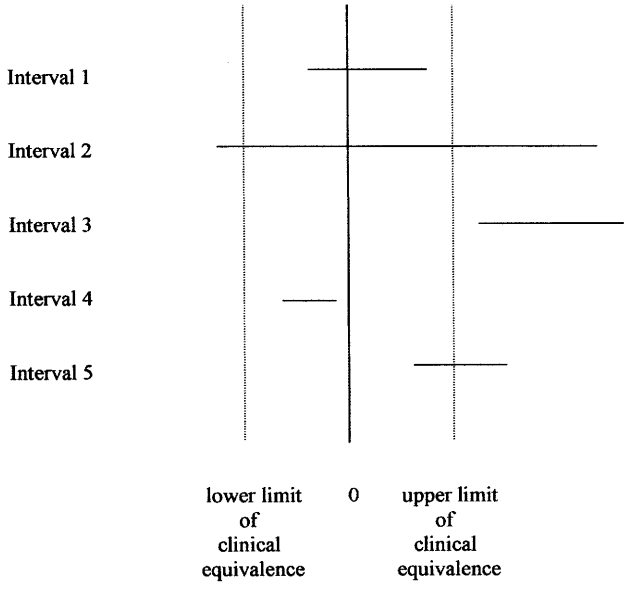

Figure 1 Clinical relevance of confidence intervals.

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

true values, would be clinically interesting. Therefore, this study shows this variable to be important.

(4) The CI does not include zero, but all values inside the upper and lower CI limits, if they were the true values, would not be clinically interesting. Therefore, this study shows this variable, while having some small effect, is not clinically important.

(5) The CI does not include zero, but only some of the values inside the upper and lower CI limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable has at least a small effect, and may be clinically important. Further study is required in order to better estimate the magnitude of this effect.

Non-significant *P*-values can be associated with both situations 1 and 2 of Figure 1, but a *P*-value alone cannot distinguish between the very different conclusions reached from interval 1 compared to interval 2. Similarly, a 'significant' *P*-value can arise from situations 3, 4, or 5, but again cannot distinguish between these very different situations. For this reason, once a confidence interval is known, the *P*-value provides little additional information, if any, while knowing a confidence interval is crucial even after calculating a *P*-value. This is the main reason for the trend away from hypothesis testing and towards confidence intervals in the medical literature and, indeed, in all fields where statistical analyses are applied. Nevertheless, since one still finds power calculations in the literature, below we will demonstrate how power calculations are performed, before going on to show how sample size calculations may more usefully be carried out using confidence interval widths.

Power calculations

Power calculations for continuous outcomes

The goal of a power calculation is to determine an appropriate sample size such that in testing the null hypothesis (H_0) with a predetermined probability of Type I error (α), the probability of a Type II error (β) is reduced to a reasonable value. For continuous outcome measures, aside from α and β, the required inputs to a power calculation are $\delta = |\mu_1 - \mu_0|$, which denotes the 'minimal clinically important difference' between the two treatments that is worthwhile to detect, and σ_0 and σ_1, the standard deviations of the outcome measure in the control and treatment groups, respectively. These standard deviations are often difficult to estimate at the planning stage of a study. If one has upper bounds for these quantities, however, these limits can be used to find a conservative sample size, in the sense that the desired power will be at least $1-\beta$ for all standard deviations equal to or less than those used. Pilot data can be very useful for estimating upper bounds for the standard errors.

Determination of the minimum clinically important difference, δ, is usually based on a combination of previous experience (such as a pilot study), published reports, and clinical experience. When multiple variables are used as outcome measurements then the sample size should be calculated for each of these variables, and the maximum sample size across these calculations can be used.

Previous clinical trials in SLE have looked at a variety of outcome measurements. Historical endpoints in randomized, controlled studies of patients with lupus nephritis have included time to end-stage renal failure and changes in immunological markers.[4] The former endpoint is clinically valid but great strides in the care of patients with SLE has markedly increased the follow-up time necessary to register clinical deterioration.[20,21] The latter endpoint is easy to measure but for common markers (DNA binding, complement levels) is of uncertain clinical relevance. The Canadian Hydroxychloroquine Study Group,[22] for example, looked at the effect of discontinuing hydroxychloroquine sulfate in patients with stable SLE and chose time to flare as the outcome measure of interest. Based on previously published reports, the authors estimated that up to 70% of patients in the placebo group would manifest worsening disease, and concluded that a 50% reduction in the rate of flares would be a clinically important difference. In another study of hydroxychloroquine in the treatment of arthropathy in SLE,[23] outcome measures included the continuous variables of subjective joint pain and both physician and patient rated disease activity (based on 5-point scales of severity) in addition to indices of joint count and joint swelling. Their data were suggestive only of some decrease in pain in the patients taking hydroxychloroquine, although the authors state that a Type II error may have been possible. Due to the small number of patients in the study, only very large ($>52\%$) differences in joint indices would have been significant (i.e. $P < 0.05$).

Given α, β, δ, σ_1 and σ_0, the required sample size, N, for a test with Type I error equal to α to have a power of $1-\beta$ is given by:

$$N = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\left(\sigma_0^2 + \sigma_1^2\right)}{\left(\mu_1 - \mu_0\right)^2} = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\left(\sigma_0^2 + \sigma_1^2\right)}{\delta^2} \quad (1)$$

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

The value $n = 2N$ is then the total number of patients required for the study, assuming equal-sized groups. The values $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are taken from normal distribution tables. For example, $Z_{1-\alpha/2} =$ 1.64, 1.96 or 2.58 for $\alpha = 0.10$, 0.05 and 0.01, respectively, and $Z_{1-\beta} = 0.84$ or 1.28 for $\beta = 0.2$ or 0.1, respectively.

Equation (1) represents the basic form used to estimate sample sizes for superiority trials of continuous outcomes. By simple algebraic rearranging of this equation, formulae may also be obtained for δ or $1-\beta$ (power) given a sample size, N, which may aid the investigator in determining the feasibility of a study if he knows in advance approximately how many patients he can expect to enroll. For example, in an anticipated 6 year study of plasmapheresis in severe lupus nephritis[2] in which 125 patients were expected to enroll over four years, and based on a mortality rate of 0.3 per year, the authors calculated a power of 88% if they were able to reduce the death rate by a factor of 2 in the treatment arm. This trial was eventually terminated based on an interim analysis showing no benefit with plasmapheresis and after calculating estimates of conditional power for detecting a significant difference had the study continued.[24]

For example, suppose one wishes to calculate the number of SLE patients with arthritis that need to be included in a study evaluating the effect of a novel analgesic (say, a cyclooxygenase-2 inhibitor) on pain. Assume that pain is reported by the patient on a visual analogue scale (VAS) from a value of 0 (no pain) to 100 (the worst pain ever experienced), and that in a pilot study with this medication patients experienced a mean decrease in their pain score of 30 points ($\delta = 30$) which was thought to be a minimally clinically relevant decrease. Suppose that the standard deviation for the decrease in pain was 20 in both the control (say, those patients on a well-known anti-inflammatory agent) and new treatment groups. The required sample size, given $\alpha = 0.05$ and $\beta = 0.10$ (where $Z_{1-\alpha/2} = 1.96$ and $Z_{1-\beta} = 1.28$, from standard tables) can be written as:

$$N = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2\ (\sigma_0^2 + \sigma_1^2)}{30^2} = \frac{(1.96 + 1.28)^2\ 2(20^2)}{900}$$
$$= 9.3 \approx 10 \text{ per group}$$

Power calculations for binary outcomes

Binary outcomes occur when the results of the study may be expressed as quantities that are either present or absent, such as the occurrence of one or more flares in a given period of time, which either occurs or does not occur in each patient in each arm of the trial. Let P_0 be the expected proportion of occurrences of the event of interest in the control group and let P_1 be the expected proportion in the treatment group. These estimates can be based on a pilot study or the best available literature. Under the null hypothesis, we assume $H_0 : P_0 = P_1 = P$, say. Unlike the normal distribution, the variance of a binomial parameter is entirely determined by the proportion of outcomes, P, so that it does not need to be separately specified. Given a Type I error α, a Type II error β, and the expected proportions in the treatment and control groups, P_0 and P_1, respectively, we can calculate the sample size as follows:

$$N = \frac{\left(Z_{1-\alpha/2}\sqrt{2\bar{P}(1-\bar{P})} + Z_{1-\beta}\sqrt{P_0(1-P_0) + P_1(1-P_1)}\right)^2}{(P_1 - P_0)^2} \quad (2)$$

where $\bar{P}$ is the average of the expected rates in the treatment and control groups, $\bar{P} = (P_0 + P_1)/2$.

For example, suppose that a pilot study suggests that a treatment may reduce progression to dialysis over 5 years in patients with lupus nephritis to 25% from a control value of 50%, i.e. $\delta = P_1 - P_0 = 25\%$ or 0.25. Note that we are establishing a definite time frame within which to calculate proportions. If we are measuring 'time-to-dialysis', then other calculation methods are applicable (see below). In this case we would calculate N, assuming $\alpha = 0.05$ and $\beta = 0.10$, as:

$$N = \frac{\left(1.96\sqrt{2(0.375)(1-0.375)} + 1.28\sqrt{0.5(0.5) + 0.25(0.75)}\right)^2}{0.25^2}$$
$$= 76.5 \approx 77 \text{ per group}$$

The formulae described above are estimates calculated by approximating a binomial distribution with the closest fitting normal distribution. This approximation is very accurate for large sample sizes. More sophisticated techniques are available for situations in which very small numbers of binary variables are involved, or in which very small proportional outcomes (e.g. < 0.05) are anticipated.[11]

Power calculations for incidence ratios

When the variable of interest is the 'time to an event', such as time to death or time to the next flare in disease, then the required sample size can be based on group differences between the incidence rates, often

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

expressed as events per person-years of risk.[25] We can let λ_0 denote the incidence rate in the control group, and let λ_1 be the incidence rate in the treatment group. For mathematical simplicity, we often assume that the distribution of times to events in each of these groups follow exponential distributions, and that the sampling distribution of mean time to an event is well approximated by a normal distribution. The latter assumption is usually reasonable for large sample sizes, but the exponential distribution, which implies a constant hazard rate, may not always be appropriate. If not, other distributions such as the Weibull may be used. Under these conditions, it can be shown that:

$$N = \frac{\left(Z_{1-\alpha/2}\sqrt{2\bar{\lambda}^2} + Z_{1-\beta}\sqrt{\lambda_1^2 + \lambda_0^2}\right)^2}{(\lambda_1 - \lambda_0)^2} \quad (3)$$

where $\bar{\lambda} = (\lambda_1 + \lambda_0)/2$

Any observation or follow-up that is terminated before the expected event has occurred is referred to as having been censored. Equation (3) does not take into account censoring, however, and refers only to the situation in which each patient is followed until the event in question has occurred. Time to event analysis with censoring is of obvious importance in superiority studies in patients with SLE where enrollment into the study is terminated at one point in time, but follow-up continues for a specific number of years. This methodology is of particular use in studies on lupus nephritis where extended follow up is required to witness progression to worsening disease and end-stage renal failure or for those studies which attempt to reduce the mean risk of a flare in patients with SLE over time. More complex statistical analysis reveals the following:[11]

$$N = \frac{\left(Z_{1-\alpha/2}\sqrt{2f(\bar{\lambda})} + Z_{1-\beta}\sqrt{f(\lambda_1) + f(\lambda_0)}\right)^2}{(\lambda_1 - \lambda_0)^2} \quad (4)$$

where

$$f(\lambda) = \lambda^3 T / \left(\lambda T_1 - e^{-\lambda(T-T_1)} + e^{-\lambda T}\right)$$

for which subjects are enrolled for T_1 years and the total duration of the study is T years.

For example, suppose that a new immunosuppressive agent is introduced to treat patients with severe, life-threatening lupus. If we estimate that the mortality rate in this select group of patients is $\lambda_0 = 0.3$ per patient-year and we are hoping to reduce this rate by a factor of 2 (such that $\lambda_1 = 0.15$) then we can calculate the required number of patients based on an anticipated 3 year enrollment and 6 year total follow-up, and assuming a power of 80% and a Type I error rate of 5%. We calculate:

$$f(\bar{\lambda}) = 0.0804$$
$$f(\lambda_0) = 0.1230$$
$$f(\lambda_1) = 0.0463$$
$$N = \frac{\left(1.96\sqrt{2(0.0804)} + 1.28\sqrt{0.0463 + 0.123}\right)^2}{(-0.15)^2}$$
$$= 76.6 \approx 77 \text{ per group}$$

Sample size based on confidence intervals

As discussed above, there has been a strong trend away from hypothesis testing and *P*-values towards the use of confidence intervals in the reporting of results from biomedical research.[26] Since the design phase of a study should be in sync with the analysis that will eventually be performed, sample size calculations should be carried out on the basis of ensuring adequate numbers for accurate estimation of important quantities that will be estimated in our study, rather than by power calculations.

The calculation of sample size via confidence interval widths often results in different sample sizes compared to power calculations, since the focus shifts from simply showing that one can reject a null hypothesis to accurately estimating treatment differences in important outcomes. The question of how accurate is 'accurate enough' (i.e. how narrow we should ensure the widths of the confidence intervals will be) can be addressed by carefully considering the results you would expect to get and making sure your confidence interval will be small enough to land in intervals numbered 1,3 or 4 of Figure 1 with high probability. The determination of an appropriate width is a non-trivial exercise, requiring careful thought about what is likely to be observed in the trial and about what is an appropriate region of clinical equivalence.

Sample size calculations for continuous outcomes

In order to estimate sample size for continuous variables, let μ_1 and μ_2 be the means of two populations being compared. Assume that we wish to estimate the difference $\mu_1 - \mu_2$ to an accuracy of a total CI width of ω, so that we will be able to report a confidence interval of the form 'estimate $\pm d$', where $d = \omega/2$. As before, let $Z_{1-\alpha/2}$ be the appropriate normal distribution quantile (for example, $Z = 1.96$ for a 95% confidence interval). Let σ_1 and σ_2 be the

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

standard deviations in the treatment and control populations, respectively, for the measure of interest. The required sample size per group is then calculated as:

$$N = \frac{Z^2_{1-\alpha/2}(\sigma_1^2 + \sigma_2^2)}{d^2} = \frac{4Z^2_{1-\alpha/2}(\sigma_1^2 + \sigma_2^2)}{\omega^2} \quad (5)$$

For example, suppose we look again at the study that compares two different analgesics on painful arthropathy in patients with SLE. Assume that the standard anti-inflammatory reduces pain by 30 points (on a 0 to 100 visual analogue scale) with a known standard deviation of 20, but that the new therapy is expected to reduce pain by a factor of 50 points with an estimated standard deviation of only 15 points. We would like to estimate this true difference between treatments to within a value of $d = 10$ and by so doing hope to detect differences of 20 points. The 95% CI will also be far enough away from 0 (at least 10 points) to derive some clinical relevance from the results (assuming our predictions of the pain reduction rates are correct). The number of patients needed per group, N, is:

$$N = \frac{Z^2_{1-\alpha 2}(\sigma_1^2 + \sigma_2^2)}{d^2} = \frac{1.96^2(20^2 + 15^2)}{10^2}$$
$$= 24.01 \approx 24 \text{ patients per group}$$

Sample size calculations for binary outcomes

Calculation of sample size for proportions proceeds in a similar fashion. Let P_1 and P_2 be the two proportions whose difference we would like to estimate to a total CI width of $\omega = 2d$; then:

$$N = \frac{Z^2_{1-\alpha/2}(P_1(1-P_1) + P_2(1-P_2))}{d^2} = \frac{4Z^2_{1-\alpha/2}(P_1(1-P_1) + P_2(1-P_2))}{\omega^2} \quad (6)$$

where N represents the required sample size for each group.

For example, suppose we would like to design a study with two types of immunosuppressive regimens to measure the difference in progression to dialysis over a period of 5 years. Assume that the standard therapy gives a $P_1 = 0.5$ (50%) rate of progression to dialysis, and that the new treatment may improve this to $P_2 = 0.25$ (25%). We would like to estimate the true difference in treatments to within $d = 0.15$ so that not only will we be able to detect the expected difference of 25%, but the 95% confidence interval will be far enough away from 0 so that we can make a more definitive conclusion as to the clinical utility of the new technique (recall Figure 1). We calculate:

$$N = \frac{Z^2_{1-\alpha/2}(P_1(1-P_1) + P_2(1-P_2))}{d^2} = \frac{1.96^2(0.5(1-0.5) + 0.25(1-0.25))}{0.15^2}$$
$$= 74.7 \approx 75 \text{ per group}$$

Discussion

Sample size calculations for several additional study designs deserve mention. The estimation of sample size in studies where the patients may serve as their own controls (such as in paired or crossover studies) is particularly useful in certain studies in SLE when total patient numbers may be limited. This method takes into account the previously estimated correlation between the two responses of an individual patient (expressed as the correlation coefficient ρ).[10,12] N is first calculated per group for the appropriate outcome variable of interest, and then the total number of patients required in a crossover study is $n = N(1-\rho)$. Such study designs reduce the total number of patients required, and are especially useful for chronic conditions such as pain or time-to-flare. Obviously, crossover designs cannot be employed for non-reversible outcomes such as dialysis or death. In addition, the individual treatments must have no carry-over effects on the patients that would bias outcomes after crossover.

Specialized methods for the calculation of sample size for the situation in which the outcome variable of interest is categorical, such as a Likert scale or a simple rating scale have been described.[27,28] Often, however these outcomes can be considered to be close enough to continuous measures so that simpler formulae can be applied.

One factor of particular relevance in studies of patients with SLE is the number of patients who fail to complete a study, especially given the length of anticipated follow-up time required to establish a reasonable incidence of certain end-points (progression to dialysis, for example). If N is a sample size calculated assuming no drop-outs, then $N_d = N/(1-D)^2$ is the sample size required in a population whose drop-out rate is expected to be D.[10] Here, patients who drop out of the study are assumed to take on the event rate of the control group once they are out of the study. If loss to follow-up is a central issue, and if it can be assumed that loss to follow-up will occur at an equal rate per group (rate L), and if there is no bias in loss to follow-up, then a simple

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011

sample-size adjustment, $N_L = N/(1-L)$, can be used. More sophisticated and exact calculation methods are available.[12,29,30]

One of the 'catch-22's of sample size calculations is that estimates need to be provided for parameters that are clearly never known before the experiment is performed (otherwise the experiment would not need to be performed!). Obviously, the sample size estimates will vary greatly depending on these unknown inputs, so that robustness to these inputs becomes a concern. A conservative sample size estimate may be derived by selecting the values of the parameters which lead to the maximum possible sample size within their feasible range. Alternatively, Bayesian sample size calculations are available which explicitly take into account the uncertainty in the inputs.[31]

Finally, while it is important for the clinician to appreciate the calculation of sample size in superiority trials and to become familiar with these estimates, sample size considerations should usually be discussed with an experienced statistician. Many factors can influence the sample size in practice, so the choice is not as simple as selecting values to plug into formulae. For example, a statistician may suggest an alternate design for answering the clinical question of interest that may be much more efficient, or easier to carry out in practice. Many sample size tables and software packages are available which offer easy calculations, given the values of the relevant parameters. Expert statistical consultations and more advanced statistical software have, for the most part, replaced the need for manual calculations on the part of the clinician, but no sample size calculations can obviate the need for a well-designed controlled clinical trial that measures standardized, clinically relevant differences in outcome. Only through such studies will clinicians further the treatment of systemic lupus erythematosus and improve the lives of patients with this disease.

Acknowledgements

Dr Joseph is a Research Scholar of the Fonds de la recherche en santé du Québec.

References

1 Carette S, Klippel JH, Decker JL *et al*. Controlled studies of oral immunosuppresive drugs in lupus nephritis. A long-term follow-up. *Ann Intern Med* 1983; **99**: 1–8.
2 Lewis EJ, Hunsicker LG, Lan S-P *et al*. A controlled trial of plasmapheresis therapy in severe lupus nephritis. *N Engl J Med* 1992; **326**: 1373–1379.
3 Gourley MF, Austin HA, Scott D *et al*. Methylprednisolone and cyclophosphamide, alone or in combination, in patients with lupus nephritis. A randomized, controlled trial. *Ann Intern Med* 1996; **125**: 549–557.
4 Balow JE. Therapeutic trials in lupus nephritis. Problems related to renal histology, monitoring of therapy and measures of outcome. *Nephron* 1981; **27**: 171–176.
5 Bellamy N, Buchanan WW, Esdaile JM *et al*. Ankylosing spondylitis antirheumatic drug trials. II. Tables for calculating sample size for clinical trials. *J Rheumtol* 1991; **18**: 1709–1715.
6 Bellamy N, Anastassiades TP, Buchanan WW *et al*. Rheumatoid arthritis antirheumatic drug trials. II. Tables for calculating sample sizes for clinical trials of antirheumatic drugs. *J Rheumtol* 1991; **18**: 1901–1907.
7 Felson DT, Anderson JJ, Boers M *et al*. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995; **38**: 727–735.
8 Strand V, Gladman D, Isenberg D *et al*. Outcome measures to be used in clinical trials in systemic lupus erythematosus. *J Rheumatol* 1999; **26**: 490–497.
9 Liang MH, Fortin PR. Viewpoint. Response criteria for clinical trials in systemic lupus erythemtosus. *Lupus* 1995; **4**: 336–338.
10 Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Conrol Clin Trials* 1981; **2**: 93–113.
11 Lemeshow S, Hosmer DW Jr, Klar J, Lwanga SK. *Adequacy of Sample Size in Health Studies*. John Wiley & Sons Ltd: Chichester, 1990.
12 Donner A. Approaches to sample size estimation in the design of clinical trials—a review. *Stat Med* 1984; **3**: 199–214.
13 Gardner MJ, Altman DG. Estimating with confidence [editorial]. *Br Med J* 1988; **296**: 1210–1211.
14 Brophy JM, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by reverend Bayes. *JAMA* 1995; **273**: 871–875.
15 Rothman KJ. Writing for Epidemiology. *Epidemiology* 1998; **9**: 333–337.
16 Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. *Stat Med* 1989; **8**: 803–811.
17 Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *Am Scien* 1988; **76**: 159–165.
18 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 ''negative'' trials. *N Engl J Med* 1978; **299**: 690–694.
19 Felson DT, Anderson J. Evidence for the superiority of immunosuppressive drugs and prednisone over prednisone alone in lupus nephritis. Results of a pooled analysis. *N Engl J Med* 1984; **311**: 1528–1533.
20 Gladman DD. Prognosis and treatment of systemic lupus erythematosus. *Curr Opin Rheumatol* 1996; **8**: 430–437.
21 Uramoto KM, Michet CJ Jr, Thumboo J *et al*. Trends in the incidence and mortality of systemic lupus erythematosus, 1950–1992. *Arthritis Rheum* 1999; **42**: 46–50.
22 The Canadian Hydroxychloroquine Study Group. A randomized study of the effect of withdrawing hydroxychloroquine sulfate in systemic lupus erythematosus. *N Engl J Med* 1991; **324**: 150–154.
23 Williams, HJ, Egger MJ, Singer JZ *et al*. Comparison of hydroxychloroquine and placebo in the treatment of the arthropathy of mild systemic lupus erythematosus. *J Rheumatol* 1994; **21**: 1457–1462.
24 Lachin JM, Lan S-P, the Lupus Nephritis Collaborative Study Group. Termination of a clinical trial with no treatment group difference: the Lupus Nephritis Collaborative Study Group. *Control Clin Trials* 1992; **13**: 62–79.
25 George SL, Desu MM. Planning the size and the duration of a clinical trial studying the time to some critical event. *J Chron Dis* 1974; **27**: 15–24.
26 Borenstein M. The case for confidence intervals in controlled clinical trials. *Control Clin Trials* 1994; **15**: 411–428.
27 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *Br Med J* 1995; **311**: 1145–1148.
28 Whitehead J. Sample size calculations for ordered categorical data. *Stat Med* 1993; **12**: 2257–2272.
29 Schork MA, Remington RD. The determination of sample size in treatment-control comparisons for chronic disease studies in which drop-out or non-adherence is a problem. *J Chron Dis* 1967; **20**: 233–239.
30 Halperin M, Rogot E, Gurian J, Ederer F. Sample sizes for medical trials with special reference to long-term therapy. *J Chron Dis* 1968; **21**: 13–24.
31 Joseph L, du Berger R, Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determinations. *Stat Med* 1997; **16**: 769–781.

Downloaded from lup.sagepub.com at MCGILL UNIVERSITY LIBRARY on December 3, 2011