ELSEVIER

Computer Methods and Programs in Biomedicine 49 (1996) 137–147

computer methods and programs in biomedicine

A computer program for sample size and power calculations in the design of multi-arm and factorial clinical trials with survival time endpoints

Ranjini Natarajan*[a], Bruce W. Turnbull[a], Elizabeth H. Slate[a], Larry C. Clark[b]

[a]*School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853-3801, USA*
[b]*University of Arizona Cancer Center, Tucson, AZ, USA*

Accepted 9 January 1996

Abstract

This paper presents a computer program for use in the design of long-term clinical trials with multiple treatment arms in which the primary outcome variables are censored survival times. The treatment arms may be structured as a one-way or multi-way factorial design. It is assumed that patients are entered and randomized to a treatment arm during an accrual period. The patients are then followed for a fixed period during which there may be dropouts. Various distributional assumptions can be used to model the survival times. These include an option in which there is an effect of treatment duly after a lag or delay time. The program then computes the power of various statistical tests of hypotheses concerning treatment differences, interactions and trends. The power computations are "exact" in that they use the Monte Carlo method to obtain Type I and II error probabilities. However the program also outputs the normal approximations for comparison, although they are typically not accurate in these situations. Fisher's LSD method is used to adjust for the multiple comparisons. By comparing the power for various sets of design parameters, such as sample size, numbers of factor levels, patient accrual rate, and length of follow-up, an appropriate design can be constructed. Two examples are provided. The first is a simple one-way layout with multiple treatment arms; the second a two-way factorial design for a proposed large scale cancer chemoprevention trial.

Keywords: Clinical trials; Power; Sample size; Multiple treatments; Factorial design; Treatment lag; Multiple comparisons; Fisher's LSD; Stratification; Monte Carlo method

1. Introduction

Until recently, factorial designs were rarely considered for the conduct of long-term randomized clinical trials. However there has been much renewed interest in the past few years, in part because such designs are natural candidates for disease prevention trials where it is desired to examine the effect of various nutritional supplements, pharmacologic agents or other potential prophylactic factors. For example, the Physicians' Health Study [1] employed a 2 × 2 design with the

* Corresponding author, Center for Statistical Science, Box G-H, Brown University, Providence, RI 02912, USA. E-mail: ranjini@stat.brown.edu.

0169-2607/96/$15.00 © 1996 Elsevier Science Ireland Ltd. All rights reserved
PII 0169-2607(96)01717-8

two factors being aspirin use and betacarotene use. A more complicated design was used in the recent Linxian study [2,3]. In this 6-year prospective intervention trial, there were four factors consisting of four vitamin/mineral groups, each at two levels-present or absent. A fractional 2^4 design was employed to investigate the effects on mortality and cancer incidence.

Of course, disease treatment trials can benefit from consideration of factorial designs also. For example, Staquet and Dalesio [4] describe a lung cancer trial in which 600 subjects were randomized into a 2×2 design to evaluate the effectiveness of chemotherapy (yes/no) and of immunotherapy (yes/no). Beck et al. [5] describe use of the same 2×2 design in the MDRD trial where the effect of two different diets and two different blood pressure controls were evaluated to investigate their effect on glomerular filtration rate in patients with chronic renal disease.

A general introduction to factorial designs for randomized clinical trials has been given by Byar and Piantadosi [6] and Byar, Herzberg and Tan [7]. Brittain and Wittes [8] studied the effect that interaction and non-compliance can have on the power to detect treatment (main) effects in a factorial design especially in comparison with the "one at a time" design. Slud [9] gives the theoretical development of semi-parametric methods of analysis of a 2×2 design, when the endpoint is survival time.

For comparing just two treatment arms with a survival endpoint, there are available a number of statistical tables, nomograms and computer programs for power and sample size calculations [10–14,16,17]. Computer programs for two treatment designs involving group sequential monitoring are also available [18–20]. Our program does not specifically address group sequential designs; however by inflating the sample size from the fixed sample design by the appropriate factor, the corresponding sample size for the group sequential test can be found [21].

Makuch and Simon [22] gave tables for designing trials with multiple arms in a one-way layout. Their method assumed exponential survival times and used normal approximations to obtain the power. Also their sample sizes were stated in terms in numbers of events observed, so further approximations are needed to convert this information into numbers of subjects and length of follow-up that would be needed. Peterson and George [23] have extended the methods of [10,22] to provide number of events requirements for testing for an interaction effect in $2 \times k$ factorial design. They used the results of [13] to obtain sample sizes and trial duration.

The purpose of this paper is to present and describe a computer program for use in planning factorial or multi-arm clinical trials. The primary outcome variable is a time to an event of interest, for example, death or onset or recurrence of disease. We assume that, after the start date of the study, there is an accrual period during which patients are recruited. After a fixed period of time from the study start date during which patients are followed, the data set is closed and a statistical analysis is performed. During this follow-up period some patients drop out and are assumed lost. Before initiating such a trial it is important to ensure that the sample sizes and follow-up period are sufficient to guarantee adequate power to detect treatment differences that are considered meaningful; also adequate power to test trend, interaction and other hypotheses of interest. The availability of such information in the planning stages of a large-scale clinical trial is extremely useful as it can indicate areas where scarce resources are best spent. Our program provides information on these issues by simulating a clinical trial according to the design specifications of the user. Exact power calculations using computed critical values are performed for the overall test of homogeneity among treatment groups, for a test of interaction and for tests for trend and other linear combinations of the group incidence rates. In addition, the user can investigate the sensitivity of the analysis to different distributional assumptions.

The methodology is described in Section 2. The computer program is described in Section 3. Some sample runs are described in Section 4. Hardware and software specifications are given in Section 5. The program is written in C and versions are currently set up to run interactively on a SUN workstation and an IBM PC or compatible.

2. Description of the methods

2.1 The model

Factorial designs permit the study of several factors simultaneously. We consider a full factorial design with multiple factors, labelled A, B,...., say. We will think of factors as different "treatments", but in fact some of these factors may actually be stratifying covariates used in the randomization. If factor A has a levels, factor B has b levels, etc., then we say we have an $a \times b \times ...$ full factorial design and there are $k = a \times b \times ...$ factor level combinations or "groups". For example, the Physicians' Health Study [1] was a 2×2 design with four groups; factor A was aspirin use (with levels aspirin or aspirin placebo), factor B was betacarotene use (with levels betacarotene or betacarotene placebo). Of course, an important special case is the one-way layout with just one factor A, at $k = a$ levels.

The outcome by which the k groups are to be compared is the time from entry into the study until the time of first occurrence of a specific event of interest, failure say. During the accrual period of V time units, patients enter at a uniform rate and are randomly assigned to one of the treatment groups. If $V = 0$, all subjects enter simultaneously. After the end of the accrual period, the study continues for a further τ (≥ 0) time units. Thus the maximum possible follow-up time for any one subject can be no more than $V + \tau$. We assume that subjects drop out of the study at a constant rate. The failure times for such subjects are considered censored. Non-failing patients surviving to the end of the study are also considered censored.

Let μ_j, F_j, f_j denote the mean, cumulative distribution and density functions, respectively, of the time to failure for subjects in group j ($1 \leq j \leq k$). We define the "incidence rate" parameter as the reciprocal of the mean, $\lambda_j = 1/\mu_j$. For exponentially distributed failure times, λ_j is the usual hazard rate. The null hypothesis of interest is $F_j = F$ for all j, that is, there is no difference among the treatment groups. The alternative hypothesis may be a general one of inhomogeneity or there may be interest in a specific type of departure from the null hypothesis such as trends in the levels of one or more of the factors, or the presence of interaction (synergism or antagonism) between the factors. These hypotheses are discussed in more detail in the following sections.

2.2. Test statistics

In this section we describe the various hypotheses of interest and the corresponding test statistics. We denote the observed incidence rate for group j ($1 \leq j \leq k$) by $\hat{\lambda}_j = d_j T_j^{-1}$, where T_j is the total exposure time of all subjects in the jth group and d_j is the number of failures in that group. The observed log incidence rates are defined as $\hat{\rho}_j = \ln \hat{\lambda}_j$. If the failure times in group j are exponentially distributed then $\hat{\lambda}_j$ is the maximum likelihood estimator of the true incidence rate λ_j and $\hat{\rho}_j$ is asymptotically normally distributed with mean $\rho_j = \ln \lambda_j$ and variance d_j^{-1} [24, page 25].

2.2.1. Overall test of homogeneity

To test the overall null hypothesis of no difference we use the following test statistic based on observed log incidence rates proposed by Makuch and Simon [22]:

$$\sum_{j=1}^{k} d_j(\hat{\rho}_j - \hat{\rho})^2 \tag{1}$$

where

$$\hat{\rho} = \left(\sum_{j=1}^{k} d_j\hat{\rho}_j\right)\left(\sum_{j=1}^{k} d_j\right)^{-1},$$

a weighted average of the $\{\hat{\rho}_j\}$. If the failure-time distributions are exponential, then Eq. 1 has an approximate chi-squared distribution with $(k - 1)$ degrees of freedom. However our program does not need to assume this is necessarily the case. The null hypothesis of homogeneity is rejected if Eq. 1 is greater than some critical value. Calculation of this critical value will be discussed in Section 2.4.

2.2.2. Overall test of interaction

An advantage of using factorial designs is the ability to test for interaction between factors. Typically an analysis of variance (ANOVA) is car-

ried out and the appropriate interaction sum of squares is used to test for the presence of interaction. Note here, however, that the numbers of events occurring in each group play the role of effective sample sizes and these will differ from group to group. Unfortunately, for unbalanced multi-way ANOVA, there is no exact way to partition the sums of squares so as to maintain orthogonality. Several approximate methods have been proposed [25–29], for example Henderson's method [26], and unweighted and weighted squares of means analysis proposed by Yates [25]. Little work appears to have been done comparing these methods, even for normally distributed data; however, based on the results of the simulation study [30], we adopted the weighted squares of means statistic, as being best approximated by a chi-square distribution.

To define this statistic we need to extend our group labelling notation. First suppose we have a two factor $a \times b$ design. We define d_{ij} and $\hat{\rho}_{ij}$ to be the number of failures and observed log incidence rate, respectively, in the group corresponding to level i of factor A and level j of factor B. The statistic used to test for presence of interaction is then given by:

$$\sum_{i=1}^{a} \sum_{j=1}^{b} d_{ij}(\hat{\rho}_{ij} - \hat{\rho})^2 - \left[b \sum_{i=1}^{a} d_{i\cdot} (\hat{\rho}_{i\cdot} - \hat{\rho})^2 + a \sum_{j=1}^{b} d_{\cdot j}(\hat{\rho}_{\cdot j} - \hat{\rho})^2 \right] \quad (2)$$

Here $\hat{\rho}_i$ is a weighted average of the $\{\hat{\rho}_{ij}\}$ given by:

$$\hat{\rho} = \left(\sum_{i=1}^{a} \sum_{j=1}^{b} d_{ij}\, \hat{\rho}_{ij} \right) \left(\sum_{i=1}^{a} \sum_{j=1}^{b} d_{ij} \right)^{-1}$$

This is the same quantity as defined after Eq. 1 but in the notation for this two-way layout. Further, $\hat{\rho}_{i\cdot}$ is the weighted average of the $\{\hat{\rho}_{ij}\}$ corresponding to level i of factor A, namely

$$\hat{\rho}_{i\cdot} = \left(\sum_{j=1}^{b} d_{ij}\, \hat{\rho}_{ij} \right) \left(\sum_{j=1}^{b} d_{ij} \right)^{-1}$$

and

$$d_{i\cdot} = b \left(\sum_{j=1}^{b} d_{ij}^{-1} \right)^{-1}$$

is the harmonic mean of the cell samples corresponding to the ith level of factor A. The quantities $\{\hat{\rho}_{\cdot j}\}$ and $\{d_{\cdot j}\}$ are defined analogously for groups corresponding to level j of factor B. If the $\{\hat{\rho}_{ij}\}$ are approximately normally distributed, as they are if the failure time distributions are exponential, then the statistic 2 has an approximate chi-squared distribution with $(a-1) \times (b-1)$ degrees of freedom. However our program does not need to assume this is necessarily the case. The null hypothesis of no interaction is rejected if Eq. 2 is greater than some critical value. Calculation of this critical value will be discussed in more detail in Section 2.4.

For higher-order factorial designs, the statistic 2 generalizes in a natural way to be defined as the total weighted sum of squares minus the sum of the main effects weighted sums of squares.

2.2.3. Tests for trends and other linear effects

It is often of interest to test various particular linear combinations of the group incidence rates. For example, in an $a \times b$ two factor design, this linear combination can be written as $\Sigma i\, \Sigma j\, c_{ij}\hat{\rho}_{ij}$ for specified constants $\{c_{ij}\}$. The corresponding test statistic is

$$\left(\sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij}\, \hat{\rho}_{ij} \right) \left(\sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij}^2\, d_{ij}^{-1} \right)^{-1/2} \quad (3)$$

The null hypothesis of homogeneity of incidence rates is rejected if the absolute value of Eq. 3 is greater than some critical value. If the rates $\{\hat{\rho}_{ij}\}$ are approximately normally distributed, then Eq. 3 has an approximate standard normal N(0,1) dis-

tribution under the null hypothesis. However our program does not need to assume this is necessarily the case.

Various choices for the $\{c_{ij}\}$ are of interest. For example, to test for presence of a linear trend in incidence rates corresponding to the levels of factor A, we may simply set $c_{ij} = i$. If we are interested in the pairwise difference between the high level of factor A, ($i = a$, say) and the control level ($i = 1$, say), we may set $c_{1j} = -1$, $c_{aj} = +1$ and $c_{ij} = 0$, $i = 2,\ldots, a-1$. Similarly, by appropriate choice of $\{c_{ij}\}$, we may test for any "single degree of freedom" interaction effect.

Tests of linear combinations are not limited to two factor designs. Analogous hypotheses of interest and corresponding test statistics can be formed for designs with any number of factors. For example, for three factor designs, the quantities in Eq. 3 would have triple subscripts; for one factor designs, formally set $b = 1$ in Eq. 3.

2.2.4. Multiple testing

In a factorial experiment there will be several hypotheses that are of interest, e.g. overall homogeneity of incidence rates, presence of interaction, trends or pairwise differences between levels in each of the factors. The problem of multiple comparisons is well known [31,32]. One way to preserve Type I error rates and protect against spurious significant results is to use Fisher's LSD method [31] as proposed by Makuch and Simon [22]. Here any linear combination or interaction test can only be found significant at level α if the overall hypothesis of homogeneity of rates is rejected using test Eq. 1 at level α. If this strategy is employed then it is ensured that the probability of falsely rejecting any true hypothesis does not exceed α. Other less conservative approaches are possible [31], but this is the most simple.

2.3. Failure-time distributions

The program presented in this paper allows the user to specify different choices for the failure-time distribution F_j. One way to characterize a failure time distribution is by its hazard rate function, $h(t)$ say. The hazard rate function is related to the cumulative distribution function (*cdf*) $F(t)$ by the relationship:

$$F(t) = 1 - \exp\{-\textstyle\int_0^t h(y)dy\} \tag{4}$$

The choice $h(t) = \alpha t^{\alpha-1}/\beta^\alpha$ corresponds to the Weibull *cdf* where $F(t) = 1 - \exp\{-(t/\beta)^\alpha\}$. Here α is termed the *shape* parameter and β the *scale* parameter. The mean and variance of the Weibull distribution are given by $\beta\Gamma(1 + \alpha^{-1})$ and $\beta^2[\Gamma(1 + 2\alpha^{-1}) - \Gamma^2(1 + \alpha^{-1})]$ where $\Gamma(\cdot)$ is the gamma function. The exponential distribution corresponds to the special case when the shape parameter $\alpha = 1$. In this case the hazard rate is constant, equal to λ, say, where $\lambda = 1/\beta$. The mean and variance are λ^{-1} and λ^{-2}, respectively.

Another commonly used failure time distribution is the lognormal, with *cdf* and density function given by $F(t) = \Phi\ ((\ln t - \mu)/\sigma)$ and $f(t) = (1/\sigma t)\phi((\ln t - \mu)/\sigma)$, respectively. Here $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal *cdf* and density, respectively. The hazard function is given by $h(t) = f(t)/[1 - F(t)]$ and the median, mean, and variance are given by μ, $\exp(\mu + \sigma^2/2)$ and $\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$, respectively.

Under the overall null hypothesis of homogeneity, the failure time distributions of all groups are the same. The program allows three possible choices — exponential, Weibull or lognormal. The powers of the various tests are computed under the alternative hypothesis. Under this hypothesis, the failure distributions of the different groups have the same form (either exponential, Weibull or lognormal), but with parameters that may vary from group to group. In addition to the above three families of distributions, two further choices are available under the alternative hypothesis. These are the lagged exponential and lagged Weibull. The lagged Weibull is defined by its hazard rate function:

$$h(t) = \begin{cases} h_0(t) \text{ if } t < t^* \\ h_1(t) \text{ if } t \geq t^* \end{cases}$$

where $h_i(t) = \alpha_i t^{\alpha_i - 1}/(\beta_i^{\alpha_i})$ for $i = 0,1$. Here t^* is a further parameter that represents the length of a lag period during which the hazard rate remains the same as that specified by the null hypothesis, i.e. $h_0(\cdot)$, but after which the treatment takes

effect, changing the hazard rate to $h_1(\cdot)$. The lagged exponential is a special case where $\alpha_0 = \alpha_1 = 1$. The provision of lagged distributions is motivated by our interest in disease prevention studies where it might be expected that intervention does not take effect immediately [33,34].

By being able to postulate different distributions under the null and alternative hypotheses, the user of our program can investigate the sensitivity of the power to a variety of distributional assumptions and thus be able to decide on a robust design.

2.4. Monte Carlo approach

In this section we will describe the simulation of a clinical trial according to the design specifications of the user. First, failure times are generated for each patient under the null as the minimum of a variate drawn from the selected failure-time distribution and a censoring time. The test statistics given in Eqs. 1 and 2 are computed using the generated data. The test statistic 3 is also computed for the linear combinations specified. This procedure is repeated N times where N is some specified (large) number. This yields N independent realizations of Eqs. 1–3. From the generated distribution of each statistic, the $100(1-\alpha)$ percentiles are calculated for $\alpha = 0.05, 0.01$. For test statistics 1 and 2 we refer to these percentiles as "exact" cut-off points (or critical values). However, since the exact distribution of Eq. 3 is not symmetric and the tests based on it are two-sided, we compute two "exact" cut-off points for it: an upper and lower $\alpha/2$ value. These are simply the upper and lower $\alpha/2$ quantiles of the generated statistic values.

For the statistics 1 and $\alpha = 0.01, 0.05$, we also record the relative frequency that the generated values exceed $\chi^2_{k-1}(\alpha)$, the upper 100α percentage point of the chi-square distribution with $k-1$ degrees of freedom. This provides a check on the accuracy of the normal approximation proposed by Makuch and Simon [22]. Similar checks are provided on the normal approximations for the statistics 2 and 3. However it should not be expected that the distribution of these statistics is approximated well by a chi-square in all circumstances.

Simulated failure times are generated similarly under the alternative hypothesis specified. For each test statistic (1–3), the proportion of generated values that lie above the exact cut-off, computed under the null as above, provide an estimate of the power of each of these tests. The power of the tests that use the normal approximation for critical values can also be computed by considering the approximate cut-off points in place of the exact ones.

3. Program description

A flow chart for the program is given in Fig. 1. Some sample runs are described in the next section. The program consists of three parts: (1) input module, (2) simulation module, (3) output module. We will now describe each of these modules in detail.

3.1. Step 1: input module

The user is first prompted for the name of the output file where the input specifications and all output is recorded. The design layout (e.g., one-way or higher-way) to be simulated is entered next. If a higher-way design is selected, the user is prompted for the number of factors. The number of

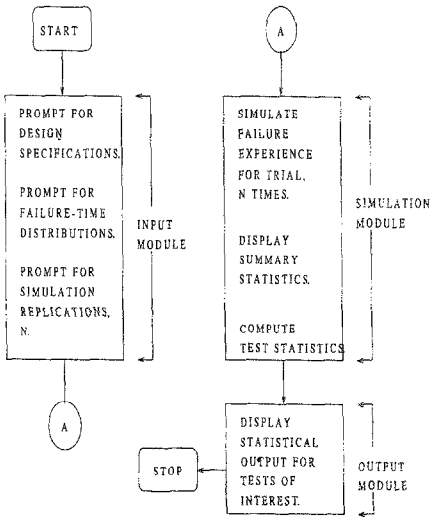

Fig. 1. Flow chart for the program

levels for each factor and the number of subjects in each treatment combination arm are entered next. The time-unit for the study (hours, days, weeks, months or years) is now inputted. This item is not actually used by the program in any of the calculations, but is useful for setting the context of

Table 1: Overall Test of Equality of Rates

The exact $100(1-\alpha)$ percentile cut-off and corresponding power are displayed for $\alpha = 0.05$ and 0.01. In addition the approximate significance level, cut-off and power using the approximate chi-