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, Arizona, USA*

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 only 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 followup, 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.

Key Words: Clinical trials, Power, Sample size, Multiple treatments, Factorial design, Treatment lag, Multiple comparisons, Fisher's LSD, Stratification, Monte Carlo method.

*Corresponding Author: Bruce W. Turnbull,
227 ETC, School of Operations Research and Industrial Engineering,
Cornell University, Ithaca, New York 14853-3801, USA.
E-mail: turnbull@orie.cornell.edu

July 31, 1995

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 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 the 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— see e.g. [10],[11], [12], [13],[14],[16],[17]. Computer programs for two treatment designs involving group sequential monitoring are also available — e.g. [18], [19], [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, Sec. 3.1].

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 followup that would be needed. This problem was further considered by Liu and Dahlberg [23]. Peterson and George [24] have extended the methods of [10],[22] to provide number of events requirements for testing for an interaction effect in a $2 \times k$ 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 followup period some patients drop out and are assumed lost. Before initiating such a trial it is important to ensure that the sample sizes and followup 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 $a \times b \times \ldots$ full factorial design and there are $k = a \times b \times \ldots$ 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 followup 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} [25, page 25].

a) 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 (1) has an approximate chi-squared distribution with $(k-1)$ degrees of freedom under the null hypothesis. However our program does not need to assume this is necessarily the case. The null hypothesis of homogeneity is rejected if (1) is greater than some critical value. Calculation of this critical value will be discussed in Section 2.4.

b) 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 carried 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 [26],[27],[28],[29], [30, p.219-220], for example Henderson's method [27], and unweighted and weighted squares of means analysis proposed by Yates [26]. Little work appears to have been done comparing these methods, even for normally distributed data; however, based on the results

of the simulation study [31], 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:

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

Here $\hat{\rho}$ 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 Equation (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 under the null hypothesis. However our program does not need to assume this is necessarily the case. The null hypothesis of no interaction is rejected if (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.

c) 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 $\sum_i \sum_j c_{ij}\hat{\rho}_{ij}$ for specified constants $\{c_{ij}\}$. The corresponding test statistic is

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

The null hypothesis of homogeneity of incidence rates is rejected if the absolute value of (3) is greater than some critical value. If the rates $\{\hat{\rho}_{ij}\}$ are approximately normally distributed, then (3) has an approximate standard normal N(0,1) distribution 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 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 (3) would have triple subscripts; for one factor designs, formally set $b = 1$ in (3).

d) 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 [32], [33]. One way to preserve Type I error rates and protect against spurious significant results is to use Fisher's LSD method [32, page 3] 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 (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 [32], 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:

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

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 $f(t) = \frac{1}{\sigma t}\phi((\ln t - \mu)/\sigma)$, 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, see e.g. [34],[35].

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 (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 (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 (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, 2, 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 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 the study. The user is then prompted for the length of the accrual period, the overall length of the study (analysis time) and the dropout rate among individuals in the study. Next, the form and parameter values of the failure-time distributions to be simulated under the null and alternative hypothesis are specified. For the exponential and Weibull distributions the mean and standard deviation of the distribution selected under the null are printed for verification purposes. The user is then prompted to enter the number (possibly 0) of linear combinations of log incidence rates to be tested. Denote this number by L_C, say. If $L_C \geq 1$, the user is further prompted to input the coefficients, in standard group order, for each of the L_C combinations, as illustrated in the sample terminal session in Appendix B. Finally, the desired number N of simulation runs (or replications) under the null and alternative are entered. Optionally, the program allows the user to specify the random number seed to initiate the simulation so that the results are repeatable.

3.2 *Step 2: Simulation Module*

The program simulates the failure experience of a clinical trial with the design specified using the input of Step 1. This is done by simulating entry, dropout and failure times under the null hypothesis for each patient with the specified sample sizes. This simulation is then replicated N times. A table of summary statistics containing the average numbers of failures generated (along with standard errors) is printed for each group. The empirical distributions of the statistics (1)—(3) are constructed and the exact and approximate cutoff points computed. The whole procedure is then repeated for failure times under the alternative hypothesis specified in Step 1.

3.3 *Step 3: Output Module*

The program prints the following results to the screen and designated output file:

- 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-square cut-off are also displayed for purposes of comparison with [22].

- Table 2: Overall test of interaction.

  This table is printed only if there are two or more factors. It contains output analogous to that in Table 1 but for the interaction statistic given by (2).

- Table 3: Two-sided tests of the individual linear combinations of log incidence rates

  For each of the L_C linear combinations specified, the exact upper and lower (denoted by U and L respectively) cut-off and power are reported. The approximate normal cut-off , significance level and power are also displayed for comparison purposes. Note that, under the normal approximation, the distribution of (3) is symmetric, so only two-sided significance levels and powers are reported for the approximate test.

- Table 4: Proportion of runs in which at least one of L_C linear combination is significant.

This table is designed to illustrate the multiple comparisons phenomenon and is displayed only if more than one linear combination is tested ($L_C \geq 2$). The proportion of simulation runs (under the null and alternative) in which at least one combination is significant is reported. The numbers along the null row reflect the inflated Type I error if each combination were tested at the unadjusted α significance level.

- Table 5: PROPORTION OF RUNS IN WHICH AT LEAST ONE OF L_C LINEAR COMBINATIONS AND OVERALL TEST OF EQUALITY IS SIGNIFICANT.

  If more than one linear combination is tested ($L_C \geq 2$), the proportion of simulation runs (both under the null and alternative) in which at least one combination and the overall test statistic is significant is reported. The numbers along the null row reflect the adjusted Type I error if Fisher's LSD multiple comparisons method is used to evaluate the significance of individual tested linear combinations.

4 Sample runs

Sample runs for two designs are described. The first is a simple one-way layout example. The second is a large clinical trial with two factors of interest. The terminal sessions are given in Appendices A and B respectively.

4.1 *One factor design with three levels*

The first example is intended to serve as a comparison with the approximate procedure of Makuch and Simon [22]. The authors determine sample size requirements for comparative clinical trials with multiple treatment groups. Failure times are assumed to follow the exponential distribution. Using approximate chi-square cut-off's, they calculate the number of failures required per group, to achieve a pre-specified power against specific alternatives. We specify a trial using the sample size they recommend for achieving a power of 0.90 for a one-way layout with three treatment arms. Since they specify sample size in terms of number of failures, we set the dropout rate to be zero and the analysis time very large (1000 years), to assure that all failures are observed. Thus all patients are followed to failure. The alternative hypothesis under consideration is one where the ratio of the largest mean failure time relative to the smallest mean failure time is two.

The sample run is displayed in Appendix A. It can be seen from the results displayed in Table 1 that the achieved power of 0.8940 (with a simulation standard error of 0.0097) is very close to the pre-specified power 0.90, thereby validating the use of the chi-square approximation by Makuch and Simon [22] in this example of exponentially distributed data.

4.2 *Two factor chemoprevention clinical trial*

We have used the program to design a large 2×3 general population chemoprevention clinical trial involving two nutritional supplements. Factor A has two levels — placebo or supplement; Factor B has three levels, placebo, low and high dose. Thus there are six treatment combination arms or groups. The objective of this trial is to determine if the two treatments under consideration have any effect on disease and mortality. Based on the resources available, the following input numbers are used. 1200 patients are assigned to each of the 6 cells. The overall length of the study is 10 years with an accrual period of 2 years. Based on pilot study data, a dropout rate of 7.5% per year is anticipated and the distribution of the time to failure under the null hypothesis is specified as exponential with a failure rate of 2%/year. The failure-time distribution under the alternative is a lagged exponential with a lag time of 2 years. Further, the failure rates under the alternative for the 6 cells are chosen such that there is no interaction in incidence rates between the factors on a multiplicative scale (additive on the log scale). The two linear combinations of interest test for the effect of the presence of each treatment. The sample run is displayed in Appendix B. The results show that such a design is probably adequate – a 5% level exact test of overall homogeneity has power of approximately 85%; the power for testing the main effect of factor A (combination 1) is 88%; for testing high and low dose of factor B versus placebo (combination 2) it is 56%. The normal approximations agree quite well again, even though, under the alternative hypothesis, the distributions are not exponential but lagged exponential. However if the powers displayed were not sufficient, the program could be rerun with larger sample sizes, Conversely, if lower powers would suffice, smaller sample sizes could be tried. Upon iteration, a suitable combination of acceptable power, numbers of factor levels and economical sample size can be obtained. The program should then be rerun under a variety of input failure distributions and parameter values to examine the sensitivity to departure to assumptions made in constructing the design. In the application which motivated this example, such considerations led to selecting only two, not three, levels for factor B.

5 Hardware and Software Specifications

The random number generator used in the simulation module applies a linear congruential method [36, p.424]. The program was written using the C language for a SUN SPARC station [37]. CPU time for the example of Section 4.1 was 2 s on a SPARC 20, 2 s on a SPARC 10, 8 s on a SPARC 2 and 20 s on a SPARC 1. The corresponding CPU times for the example of Section 4.2 were 239, 287, 781 and 1873 s. A Microsoft C [38] program has also been compiled to run on a 386 or higher IBM PC or compatible.

6 Mode of Availability of the Program

Copies of the program are available upon request from the corresponding author, Bruce W. Turnbull.

7 Acknowledgement

This research was supported by grants from the U.S. National Institutes of Health.

Appendix A. Sample terminal session –Oneway layout

```
      ================================================================
      POWER CALCULATIONS FOR USE IN THE DESIGN OF A MULTI-ARM AND FACTORIAL
      CLINICAL TRIAL USING THE MAKUCH AND SIMON TEST STATISTIC
      ================================================================
SPECIFY FILE FOR PRINTED OUTPUT: oneway.dat
ENTER THE DESIGN TO BE SIMULATED:
                1. ONE-WAY LAYOUT
                2. HIGHER-WAY FACTORIAL DESIGN
ENTER SELECTION HERE: 1
ENTER THE NUMBER OF TREATMENT GROUPS ( > 1): 3
ARE THERE EQUAL NUMBER OF SUBJECTS IN EACH GROUP? (Y/N) y
ENTER THE NUMBER OF SUBJECTS IN EACH GROUP: 53
ENTER THE TIME UNIT? (H=HOUR, D=DAY, W=WEEK, M=MONTH, Y=YEAR) y
ENTER LENGTH OF ACCRUAL PERIOD (IN YEARS; 0 FOR SIMULTANEOUS ENTRY): 0
ENTER OVERALL LENGTH OF STUDY (IN YEARS): 1000
ENTER THE DROP-OUT RATE IN PERCENT PER YEAR: 0
WHAT IS THE UNDERLYING DISTRIBUTION OF THE TIME TO FAILURE TO BE SIMULATED?

UNDER THE NULL:
===============
                1. EXPONENTIAL
                2. LOG NORMAL
                3. WEIBULL

ENTER SELECTION HERE: 1

UNDER THE ALTERNATIVE:
======================
                1. EXPONENTIAL
                2. LOG NORMAL
                3. WEIBULL
                4. LAGGED EXPONENTIAL
                5. LAGGED WEIBULL

ENTER SELECTION HERE: 1

NULL HYPOTHESIS
===============
     ENTER FAILURE RATE IN PERCENT PER YEAR: 5
     THE MEAN TIME TO FAILURE IS          20.00 YEARS
     THE STD. DEV. OF TIME TO FAILURE IS  20.00  YEARS

ALTERNATIVE HYPOTHESIS
======================
     ENTER FAILURE RATE IN PERCENT PER YEAR:
               Group 1:        5
               Group 2:        2.5
               Group 3:        3.5
```

```
ENTER NUMBER OF LINEAR COMBINATIONS OF THE LOG MEAN RESPONSE TIME: 0
ENTER NUMBER OF SIMULATION RUNS FOR THE NULL: 1000
ENTER NUMBER OF SIMULATION RUNS FOR THE ALTERNATIVE: 1000
ENTER RANDOM NUMBER SEED (0 FOR RANDOM SEED): 2948239487

GENERATING DATA UNDER THE NULL.........PLEASE WAIT

                  AVERAGE NUMBER OF FAILURES
                  ==========================
                                              MEAN    S. E
        Group 1:                             53.00   (0.00)
        Group 2:                             53.00   (0.00)
        Group 3:                             53.00   (0.00)

GENERATING DATA UNDER THE ALTERNATIVE.........PLEASE WAIT

                  AVERAGE NUMBER OF FAILURES
                  ==========================
                                              MEAN    S. E
        Group 1:                             53.00   (0.00)
        Group 2:                             53.00   (0.00)
        Group 3:                             53.00   (0.00)

                       STATISTICAL OUTPUT
                       ==================
                 TABLE 1: OVERALL TEST OF EQUALITY OF RATES
                 ==========================================

                       NOMINAL CUT-OFF  ACHIEVED        ACHIEVED
                       SIZE    VALUE    SIG LEVEL       POWER
                       ======= =======  =========       ========
Approximate ChiSquare  0.0500  5.9915   0.0590 (0.0075) 0.9070 (0.0092)
                       0.0100  9.2103   0.0050 (0.0022) 0.7700 (0.0133)
                       ================================================
Exact (via Simulation) 0.0500  6.2601   0.0500          0.8940 (0.0097)
                       0.0100  8.7719   0.0100          0.7850 (0.0130)
```

Appendix B. Sample terminal session – Clinical trial data

```
      ================================================================
      POWER CALCULATIONS FOR USE IN THE DESIGN OF A MULTI-ARM AND FACTORIAL
      CLINICAL TRIAL USING THE MAKUCH AND SIMON TEST STATISTIC
      ================================================================
SPECIFY FILE FOR PRINTED OUTPUT: clinical.dat
ENTER THE DESIGN TO BE SIMULATED:
                1. ONE-WAY LAYOUT
                2. HIGHER-WAY FACTORIAL DESIGN
ENTER SELECTION HERE: 2
ENTER THE NUMBER OF FACTORS: 2
ENTER THE NUMBER OF LEVELS OF FACTOR 1: 2
ENTER THE NUMBER OF LEVELS OF FACTOR 2: 3
  THE SELECTED DESIGN HAS 6 TREATMENT COMBINATION GROUPS
ARE THERE EQUAL NUMBER OF SUBJECTS IN EACH GROUP? (Y/N) y
ENTER THE NUMBER OF SUBJECTS IN EACH GROUP: 1200
ENTER THE TIME UNIT? (H=HOUR, D=DAY, W=WEEK, M=MONTH, Y=YEAR) y
ENTER LENGTH OF ACCRUAL PERIOD (IN YEARS; 0 FOR SIMULTANEOUS ENTRY): 2
ENTER OVERALL LENGTH OF STUDY (IN YEARS): 10
ENTER THE DROP-OUT RATE IN PERCENT PER YEAR: 7.5
WHAT IS THE UNDERLYING DISTRIBUTION OF THE TIME TO FAILURE TO BE SIMULATED?

UNDER THE NULL:
===============
                1. EXPONENTIAL
                2. LOG NORMAL
                3. WEIBULL
ENTER SELECTION HERE: 1

UNDER THE ALTERNATIVE:
======================
                1. EXPONENTIAL
                2. LOG NORMAL
                3. WEIBULL
                4. LAGGED EXPONENTIAL
                5. LAGGED WEIBULL
ENTER SELECTION HERE: 4

NULL HYPOTHESIS
===============
     ENTER FAILURE RATE IN PERCENT PER YEAR: 2
     THE MEAN TIME TO FAILURE IS          50.00 YEARS
     THE STD. DEV. OF TIME TO FAILURE IS  50.00  YEARS
```

```
ALTERNATIVE HYPOTHESIS
======================
     ENTER LAG (IN YEARS) UNTIL TREATMENT TAKES EFFECT (0 FOR NO LAG):  2
     ENTER FAILURE RATE IN PERCENT PER YEAR (EFFECTIVE BEFORE 2.00 YEARS):  2
     ENTER FAILURE RATE IN PERCENT PER YEAR (EFFECTIVE AFTER 2.00 YEARS):
               GROUP 1:( 1, 1 ) 2
               GROUP 2:( 1, 2 ) 1.6816
               GROUP 3:( 1, 3 ) 1.416
               GROUP 4:( 2, 1 ) 1.416
               GROUP 5:( 2, 2 ) 1.1256
               GROUP 6:( 2, 3 ) 1

ENTER NUMBER OF LINEAR COMBINATIONS OF THE LOG MEAN RESPONSE TIME: 2
        ENTER THE COEFFICIENTS OF COMB 1 (IN GROUP ORDER): -1 -1 -1 1 1 1
        ENTER THE COEFFICIENTS OF COMB 2 (IN GROUP ORDER): -1 1 1 -1 1 1
ENTER NUMBER OF SIMULATION RUNS FOR THE NULL: 1000
ENTER NUMBER OF SIMULATION RUNS FOR THE ALTERNATIVE: 1000
ENTER RANDOM NUMBER SEED (0 FOR RANDOM SEED): 9287925

GENERATING DATA UNDER THE NULL.........PLEASE WAIT

                   AVERAGE NUMBER OF FAILURES
                   ==========================
                                                MEAN    S. E
                GROUP 1:( 1, 1 )               144.93  (0.36)
                GROUP 2:( 1, 2 )               145.17  (0.37)
                GROUP 3:( 1, 3 )               144.63  (0.36)
                GROUP 4:( 2, 1 )               144.56  (0.36)
                GROUP 5:( 2, 2 )               144.37  (0.35)
                GROUP 6:( 2, 3 )               145.00  (0.36)

GENERATING DATA UNDER THE ALTERNATIVE.........PLEASE WAIT

                   AVERAGE NUMBER OF FAILURES
                   ==========================
                                                MEAN    S. E
                GROUP 1:( 1, 1 )               145.13  (0.37)
                GROUP 2:( 1, 2 )               129.56  (0.35)
                GROUP 3:( 1, 3 )               117.03  (0.33)
                GROUP 4:( 2, 1 )               116.74  (0.32)
                GROUP 5:( 2, 2 )               102.30  (0.30)
                GROUP 6:( 2, 3 )                95.87  (0.30)
```

STATISTICAL OUTPUT
==================

TABLE 1: OVERALL TEST OF EQUALITY OF RATES
==

	NOMINAL SIZE	CUT-OFF VALUE	ACHIEVED SIG LEVEL	ACHIEVED POWER
Approximate ChiSquare	0.0500	11.0705	0.0610 (0.0076)	0.8620 (0.0109)
	0.0100	15.0863	0.0190 (0.0043)	0.6710 (0.0149)
Exact (via Simulation)	0.0500	11.7844	0.0500	0.8450 (0.0114)
	0.0100	16.2996	0.0100	0.6010 (0.0155)

TABLE 2: OVERALL TEST OF INTERACTION
====================================

	NOMINAL SIZE	CUT-OFF VALUE	ACHIEVED SIG LEVEL	ACHIEVED POWER
Approximate ChiSquare	0.0500	5.9915	0.0530 (0.0071)	0.0620 (0.0076)
	0.0100	9.2103	0.0080 (0.0028)	0.0150 (0.0038)
Exact (via Simulation)	0.0500	6.0300	0.0500	0.0620 (0.0076)
	0.0100	8.3564	0.0100	0.0190 (0.0043)

TABLE 3: 2-SIDED TESTS OF THE INDIVIDUAL LINEAR COMBINATION(S) OF LOG(MEAN) RESPONSE TIME
==

COMBINATION 1:
==============
-1.00 -1.00 -1.00 1.00 1.00 1.00

	NOMINAL SIZE	CUT-OFF VALUE	ACHIEVED SIG LEVEL	ACHIEVED POWER
Approximate Normal	0.0500	1.9600	0.0420 (0.0063)	0.8490 (0.0113)
	0.0100	2.5758	0.0080 (0.0028)	0.6620 (0.0150)
Exact (via Simulation)	0.0250L	-1.9446	0.0250	0.0000 (0.0000)
	0.0250U	1.8471	0.0250	0.8800 (0.0103)
			------	------
			0.0500	0.8800
	0.0050L	-2.5752	0.0050	0.0000 (0.0000)
	0.0050U	2.5528	0.0050	0.6710 (0.0149)
			------	------
			0.0100	0.6710

COMBINATION 2:
==============
-1.00 1.00 1.00 -1.00 1.00 1.00

	NOMINAL SIZE	CUT-OFF VALUE	ACHIEVED SIG LEVEL	ACHIEVED POWER
Approximate Normal	0.0500	1.9600	0.0590 (0.0075)	0.6000 (0.0155)
	0.0100	2.5758	0.0120 (0.0034)	0.3590 (0.0152)
Exact (via Simulation)	0.0250L	-2.0527	0.0250	0.0000 (0.0000)
	0.0250U	2.0352	0.0250	0.5590 (0.0157)
			0.0500	0.5590
	0.0050L	-2.6361	0.0050	0.0000 (0.0000)
	0.0050U	2.9104	0.0050	0.2500 (0.0137)
			0.0100	0.2500

TABLE 4: PROPORTION OF RUNS IN WHICH AT LEAST ONE OF 2 LINEAR COMBINATION(S) IS SIGNIFICANT (USING EXACT VIA SIMULATION CUT-OFF)

	0.0500	0.0100
Null:	0.0950 (0.0093)	0.0200 (0.0044)
Alternative:	0.9500 (0.0069)	0.7510 (0.0137)

TABLE 5: PROPORTION OF RUNS IN WHICH AT LEAST ONE OF 2 LINEAR COMBINATION(S) AND OVERALL TEST OF EQUALITY IS SIGNIFICANT (USING EXACT VIA SIMULATION CUT-OFF)

	0.0500	0.0100
Null:	0.0270 (0.0051)	0.0030 (0.0017)
Alternative:	0.8350 (0.0117)	0.5710 (0.0157)

References

[1] M.J. Stampfer, J.E. Buring, W. Willett, B. Rosner, K. Eberlein and C.H. Hennekens, The 2×2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians. Statistics in Medicine 4 (1985) 111-116.

[2] W.J. Blot, J. Li, P.R. Taylor, W. Guo, S. Dawsey, G. Wang, C.S. Yang, S. Zheng, M. Gail, G. Li, Y. Yu, B. Liu, J. Tangrea, Y. Sun, F. Liu, J.F. Fraumeni, Y. Zhang and B. Li, Nutrition intervention trials in Linxian, China: supplementations with specific vitamin/mineral combinations, cancer incidence, and disease specific mortality in the general population, J. Nat. Cancer Inst. 85 (1993) 1483-1492.

[3] J. Li, P.R. Taylor, B. Li, S. Dawsey, G. Wang, A.G. Ershow, W. Guo, S. Liu, C.S. Yang, Q. Shen, W. Wang, S.D. Mark, X.N. Zou, P. Greenwald, Y. Wu and W.J. Blot, Nutrition intervention trials in Linxian, China: multiple vitamin/mineral supplementation, cancer incidence, and disease specific mortality among adults with esophageal dysplasia, J. Nat. Cancer Inst. 85 (1993) 1492-1498.

[4] M. Staquet and O. Dalesio, Designs for Phase III trials. In Cancer Clinical Trials: Methods and Practice (M.E.Buyse, M.J.Staquet and R.J.Sylvester, Eds.) (Oxford University Press, Oxford U.K, 1984).

[5] G.J. Beck, R.L. Berg, C.H. Coggins, J.H. Gassman, L.G. Hunsicker, M.D. Schluchter and G.W. Williams, Design and statistical issues of the modification of diet in renal disease trial, Contr. Clin. Trials (1991) 12, 566-586.

[6] D.P. Byar and S. Piantadosi, Factorial designs for randomized clinical trials, Cancer Treatment Reports 69 (1985) 1055-1063.

[7] D.P. Byar, A.M. Herzberg and W-Y. Tan, Incomplete Factorial Designs for Randomized Clinical Trials, Statistics in Medicine (1993) 12, 1629-1641.

[8] E. Brittain and J. Wittes, Factorial designs in clinical trials: the effects of non-compliance and subadditivity, Statistics in Medicine 8 (1989) 161-171

[9] E.V. Slud, Analysis of factorial survival experiments, Biometrics 50 (1994) 25-38.

[10] S.L. George and M.M. Desu, Planning the size and duration of a clinical trial studying the time to some critical event, J. Chron. Dis. 27 (1974) 15-24.

[11] D. Bernstein and S.W. Lagakos, Sample size and power determination for stratified clinical trials. J. Statist. Comput. Simul. 8 (1978) 65-73.

[12] M. Wu, M. Fisher and D. DeMets, Sample sizes for long-term medical trials with time-dependent dropout and event rates, Contr. Clin. Trials 1 (1980) 111-123.

[13] L.V. Rubinstein, M.H. Gail, and T.J. Santner, Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. J. Chron. Dis. 34 (1981) 469-479.

[14] D.A. Schoenfeld and J.R. Richter, Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. Biometrics, 38 (1982) 163-170.

[15] W.H. Wooding, Planning Pharmaceutical Clinical Trials (Wiley, New York, NY, 1993).

[16] B.W. Brown, A. Chan. D. Gutierez, J. Herson, J. Lovato and J. Polsley, STPLAN 3.0: Calculations for Sample Sizes and Related Problems. Dept. of Biomathematics, M.D. Anderson Cancer Center, Univ. of Texas (1990) .

[17] Information Management Services, Inc., Interactive Power Program, Rockville, MD (1994).

[18] J. Halpern and B.W. Brown, A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. Contr. Clin. Trials, 14 (1993) 109-122.

[19] EaSt – A Software Package for the Design and Interim Monitoring of Group Sequential Trials, Cytel Software Corp., Cambridge, MA (1992).

[20] J. Whitehead and H. Brunier, PEST2.0 Operating Manual. Reading University, U.K. (1989).

[21] C. Jennison and B.W. Turnbull, Interim analyses: the repeated confidence interval approach, J. Roy. Statist. Soc. B51 (1989) 305-361.

[22] R.W. Makuch and R.M. Simon, Sample size requirements for comparing time-to-failure among k treatment groups, J. Chron. Dis. 35 (1982) 861-867.

[23] P.Y. Liu and S. Dahlberg, Design and analysis for multi-arm trials with survival endpoints, Contr. Clin. Trials 16 (1995) 119-130.

[24] B. Peterson and S.L. George, Sample Size requirements and length of study for testing interaction in a $2 \times k$ factorial design when time-to-failure is the outcome, Contr. Clin. Trials 14 (1993) 511-522.

[25] R.G. Miller, Survival Analysis, (Wiley, New York, 1981).

[26] F. Yates, The analysis of multiple classifications with unequal numbers in the different classes, J. Am. Statist. Assoc. 29 (1934) 51-66.

[27] C.R. Henderson, Estimation of variance and covariance components, Biometrics 9 (1953) 226-252.

[28] N.O. Rankin, The harmonic mean method for one-way and two-way analyses of variance, Biometrika 61 (1974) 117-122.

[29] S.R. Searle, Linear Models for Unbalanced Data, (Wiley, New York, 1987).

[30] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, (Wiley, New York, NY, 1992).

[31] E.H. Slate, R. Natarajan, B.W. Turnbull and L. C. Clark, Design of Factorial Clinical Trials, Proceedings of the 27th Symposium on the Interface, Ed: M. Meyer, Pittsburgh PA June 21–25, 1995.

[32] Y. Hochberg and A.C. Tamhane, Multiple Comparison Procedures, (Wiley, New York, NY, 1987).

[33] P.H. Westfall and S.S. Young, Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment, (Wiley, New York, NY, 1993).

[34] D.M. Zucker and E. Lakatos, Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment, Biometrika 77 (1990) 853-864.

[35] X. Luo, B.W. Turnbull, H. Cai and L.C. Clark, Regression for censored survival data with lag effects, Commun. Statist. – Theor. Meth. 23 (1994) 3417-3438.

[36] A.M. Law and D. Kelton, Simulation Modeling and Analysis, (McGraw-Hill, New York, NY, 1991).

[37] Sun Microsystems, Inc., Mountain View, California.

[38] Microsoft C, Microsoft Corp. (Redmond, Washington).

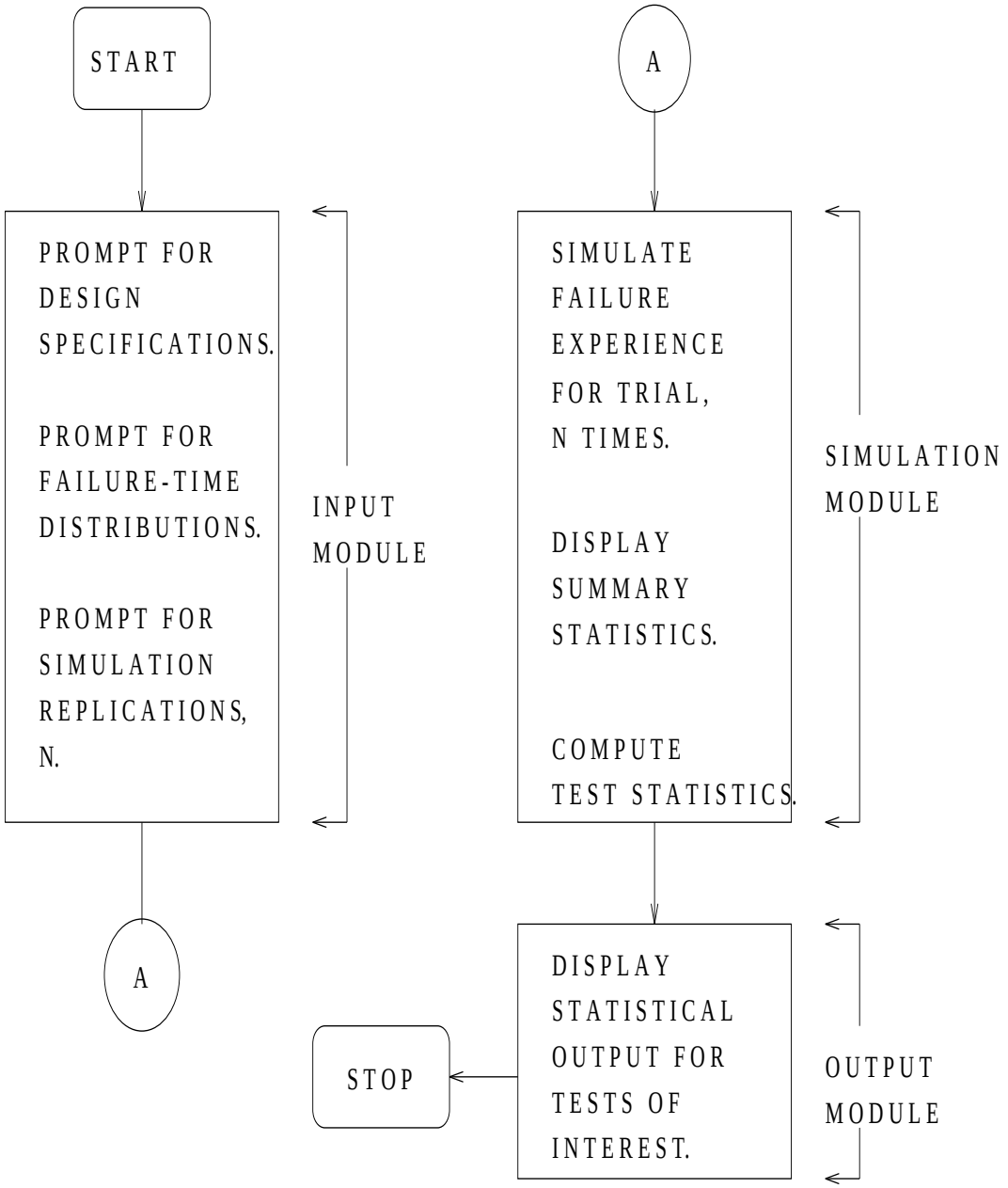

Figure 1: Flow Chart for the Program