APPLICATION OF THE NEGATIVE MULTINOMIAL DISTRIBUTION TO COMPARATIVE POISSON CLINICAL TRIALS OF MULTIPLE EXPERIMENTAL TREATMENTS VERSUS A SINGLE CONTROL

By

JOSEPH A. CHIARAPPA

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Donald R. Hoover

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2019

ABSTRACT OF THE DISSERTATION

Application of the Negative Multinomial Distribution to Comparative Poisson Clinical Trials of Multiple Experimental Treatments Versus a Single Control

By JOSEPH A. CHIARAPPA

Dissertation Director:
Donald R. Hoover

Clinical trials that compare one or more experimental treatments to a control treatment in which event incidence (i.e. incidence of disease or an adverse event) is rare often assume that comparative Poisson methodology is appropriate for modeling the number of events that occur in each treatment group. Clinical studies of multiple Poisson parameters may be conducted under one of two designs: (A) wait until a total number of events occur among all treatment groups before stopping the study, or (B) wait until a specified amount of time has passed before terminating the study. Exact tests under these approaches are based on the multinomial distribution.

In this dissertation, we consider an alternative approach termed "Design C", which is to wait until the control group accumulates a pre-specified number of events before stopping the study. The joint distribution of the number of events in the experimental treatment groups at the time of study stoppage, conditional on the number of events observed in the control group, follows a negative multinomial distribution (NMD). The minimum (respectively, maximum) number of events among the experimental treatment arms will be shown to be an appropriate test statistic for

determining whether one or more of the experimental treatments is superior (respectively, inferior) to the control at a given one-sided overall Type I error; as such, we first determine the distribution of the order statistics of the NMD. We subsequently provide tables of trial design parameters for select values of one-sided overall Type I error and pointwise power and assuming equal allocation of study subjects to the treatment groups. These studies can be improved by applying curtailed stoppage rules; that is, follow-up of the treatment arms can be discontinued prior to the control group reaching its planned number of events once the ultimate decision is known for each arm. Curtailment has substantial practical implications as reduced follow-up implies reduced study costs and more rapid knowledge of the trial results. We provide simple algorithms to estimate the expected amount of subject follow up (presented in terms of person years) that would be needed until trial termination under both uncurtailed and curtailed stopping rules. Finally, we combine the superiority and inferiority test procedures to provide a two-sided test and briefly consider pairwise comparison of the experimental treatments to each other under the Design C framework.

Acknowledgements

I am sincerely grateful for all of the support, encouragement, and advice I have received from my advisor, Dr. Donald R. Hoover. His contributions, patience, and guidance were essential for the completion of this dissertation, as well as for my academic and professional growth.

I would like to thank Dr. Ying Hung, Dr. John Kolassa, and Dr. Mark Li for serving on my dissertation committee, as well as all of the faculty members in the department of Statistics and Biostatistics for their constant willingness to share their expertise.

Finally, I want to thank my family for their unwavering love and support.

Dedication

This dissertation is dedicated to my family

Table of Contents

List of Tables

List of Illustrations

SECTION 0: INTRODUCTION

Clinical studies that compare multiple experimental treatments to a single control treatment (which may be either a placebo or a standard of care treatment) typically occur early in the drug process. For example, during drug discovery, an agent may be modified to produce several related drugs in the same drug class. It is then of interest to determine which of the derivative drugs is most effective in preventing an illness or curing a disease and which are ineffective or harmful. In phase 1 trials, various doses of a drug or vaccine may be administered to study participants with the intent of determining the minimal effective and maximum tolerable doses. In either case, a control treatment may be administered for comparison. The landscape of clinical trials has historically been a rapidly expanding one, and late-phase clinical trials which compare several experimental treatments to a single control are now more commonplace. Parmar, Carpenter, and Sydes (2014) report that approximately 20% of superiority randomized controlled trials registered between January, 2010 and July, 2012 had three or more treatment arms.

Large-scale controlled clinical trials typically have two objectives: prove a new treatment is superior in efficacy to the control and prove the new treatment is safe for consumption. In some settings, the primary outcome of the trial is a rare binomial event, such as prevention of HIV transmission, or a Poisson outcome, such as in vaccine trials. In either case, the Poisson distribution may be used as the basis for statistical comparison of the rates of events in the treatment groups; trials having the Poisson distribution as the basis for statistical inference are referred to as comparative Poisson trials (Gail, 1974).

When there are only two treatments under study, two popular methods for conducting comparative Poisson trials are described in Gail (1974). Under “Design A”,

the trial is conducted until a pre-specified total number of events among the two treatment groups are observed, while under "Design B", the trial is stopped after a pre-specified amount of time. Testing the efficacy or safety of the treatments under these two designs is based on the binomial distribution. In this dissertation, which extends consideration to comparing multiple treatments to a single control, our primary interest is in "Design C", which, to our knowledge, was first proposed by Hsu (2010). Under Design C, the study continues until a pre-specified number of events are observed in the control group; when one treatment is compared to a control under Design C, testing is conducted via a negative binomial distribution. When more than two treatment groups are under study, Designs A and B naturally extend to testing based on the multinomial distribution, while extension of Design C leads to tests based on the negative multinomial distribution (NMD).

Comparative Poisson trials of multiple experimental treatments versus a single control treatment conducted under Design C methodology will be our primary focus in this dissertation. As such, properties of the negative multinomial distribution provide the basis for hypothesis tests concerning the superiority or inferiority of experimental treatments in relation to a control. Based on these objectives, the structure of this dissertation is as follows: in Section 1, we provide the characterization and probability mass function of the negative multinomial distribution and define "balanced" and "unbalanced" NMDs. A historical overview of the NMD and examples of its practical application are also provided. In Section 2, we discuss the comparative Poisson process and its relationship to the NMD. This relationship is utilized in Section 3 to derive the order statistics of the balanced NMD. Simulation is used to compute the order statistics of

unbalanced negative multinomial distributions. The preliminary results of Sections 1, 2, and 3 are applied in Section 4, where Design C methodology is developed for clinical trials in which one or more experimental treatments are investigated for superiority to a control treatment under the assumption of equal allocation of study subjects to the trial arms. We present the main hypotheses of interest, derive an appropriate test statistic, provide tables of trial design parameters for specified combinations of overall one-sided Type I error and power, compare our results to those obtained using the Bonferroni procedure to control for multiple comparisons, and discuss the expected number of person years of follow-up until trial termination under uncurtailed and curtailed stoppage rules. In Section 5, we extend the methodology to accommodate trials which investigate treatment inferiority, combine the superiority and inferiority results into a two-sided test, and discuss pairwise comparisons of the experimental treatments to each other. Section 6 concludes with future directions implicated by the work in this dissertation.

SECTION 1: THE NEGATIVE MULTINOMIAL DISTRIBUTION

1.1: Characterization and Probability Mass Function of the Negative Multinomial Distribution

It is well known that if X denotes the number of successes in m $Bernoulli(p)$ trials, then X has a binomial distribution with parameters m and p (i.e. $X \sim Bin(m, p)$), and the probability mass function of X is given by

$$P(X = x) = \binom{m}{x} p^x (1-p)^{m-x}, x = 0, 1, 2, \dots, m$$

Now, consider the random variable Y denoting the number of failures that occur before r successes are obtained in a sequence of $Bernoulli(p)$ trials. The random variable Y has a negative binomial distribution with parameters r and p (i.e. $Y \sim NB(r, p)$), and the probability mass function of Y is given by

$$P(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y, y = 0,1,2, \dots \qquad (1.1.1)$$

Both the binomial and negative binomial distributions are predicated upon the fact that in a Bernoulli trial there are only two possible outcomes, generally referred to as "success" and "failure". Suppose instead that we conduct an experiment in which we observe *m* independent trials for which there are *n* mutually exclusive and exhaustive possible outcomes for each trial. Denote the probability of the i^{th} possible outcome by p_i, $0 \leq p_i \leq 1$ for $i = 1,2, \dots, n$, and let X_i count the number of times that the i^{th} outcome occurs in the m trials. The vector $(X_1, X_2, \dots, X_n)$ has a multinomial distribution with parameters m and $p_1, p_2, \dots, p_n$ (i.e. $(X_1, \dots, X_n) \sim M(m, p_1, \dots, p_n)$) and has probability mass function

$$P(X_1 = x_1, \dots, X_n = x_n) = m! \prod_{i=1}^{n} \frac{p_i^{x_i}}{x_i!} = \frac{m!}{x_1!\, x_2! \cdots x_n!} \prod_{i=1}^{n} p_i^{x_i} = \binom{m}{x_1, x_2, \dots, x_n} \prod_{i=1}^{n} p_i^{x_i}$$

where $x_i \geq 0$ for $i = 1,2, \dots, n, \sum_{i=1}^{n} p_i = 1$, and $\sum_{i=1}^{n} x_i = m$.

Note that since $\sum_{i=1}^{n} p_i = 1$, in theory we need only specify $n - 1$ of the probabilities p_i as parameters to characterize the distribution. However, all n probabilities are typically specified as parameters of the multinomial distribution throughout the literature, and so we will maintain this standard notation when discussing the multinomial distribution (Johnson, Kotz, and Balakrishnan, 1997, pages 31-33).

From the characterization of the multinomial distribution, it is clear that the multinomial distribution is the multivariate analogue of the binomial distribution. That is, the multinomial distribution extends the binomial distribution to $n > 2$ possible outcomes for each trial. Thus, a natural question is whether there exists an $n > 2$ outcome multivariate analogue of the negative binomial distribution whose relationship to the multinomial distribution mirrors the relationship between the negative binomial and binomial distributions. Such a distribution, the negative multinomial distribution (NMD), is characterized in Le Gall (2006) as follows: suppose (substituting $h + 1$ for n to allow outcome 0 to be the "reference" outcome as will be described below) there are $n = h + 1$ mutually exclusive and exhaustive outcomes denoted by $E_0, E_1, \dots, E_h$, which occur with probabilities $p_0, p_1, \dots, p_h$, respectively. If independent trials are conducted until the "reference" outcome E_0 occurs ν times ($\nu > 0$), then the number of occurrences $Y_1, Y_2, \dots, Y_h$ of outcomes $E_1, E_2, \dots, E_h$, respectively, during these trials will have a negative multinomial distribution with parameters ν and $p_0, p_1, p_2, \dots, p_h$ (i.e. $(Y_1, \dots, Y_h) \sim NM(\nu, p_0, p_1, \dots, p_h)$). The probability mass function is given by

$$P(Y_1, \dots, Y_h) = \Gamma\left(\nu + \sum_{i=1}^{h} y_i\right) \frac{p_0^{\nu}}{\Gamma(\nu)} \prod_{i=1}^{h} \frac{p_i^{y_i}}{y_i!}, y_j = 0,1,2 \dots, j = 1,2, \dots, h \quad (1.1.2)$$

where y_j is an observed value of the random variable Y_j for $j = 1, \dots, h$, and $\sum_{i=0}^{h} p_i = 1$. When ν is a positive integer value, the distribution is sometimes referred to as the h-variate Pascal distribution, and the probability mass function can be written as

$$P(Y_1, \dots, Y_h) = \left(\nu - 1 + \sum_{i=1}^{h} y_i\right)! \frac{p_0^{\nu}}{(\nu-1)!} \prod_{i=1}^{h} \frac{p_i^{y_i}}{y_i!}$$

When there are only two possible outcomes, E_0 and E_1, and ν is a positive integer, the probability mass function is

$$P(Y_1) = \Gamma(\nu + y_1) \frac{p_0^{\nu}}{\Gamma(\nu)} \frac{p_1^{y_1}}{y_1!} = \frac{(\nu + y_1 - 1)!}{(\nu-1)!\, y_1!} p_0^{\nu} p_1^{y_1} = \binom{\nu + y_1 - 1}{y_1} p_0^{\nu} (1 - p_0)^{y_1}$$

which, by Equation 1.1.1, is the probability mass function of a $NB(\nu, p_0)$ distribution.

Note that p_0 need not be included as a parameter in specifying the NMD. Since the outcomes $E_0, \dots, E_h$ are exhaustive, we know that $\sum_{i=0}^{h} p_i = 1$, and so when the values of $p_1, \dots, p_h$ are known, the value of p_0 is determined by $1 - \sum_{i=1}^{h} p_i$. In this dissertation, however, we will specify p_0 as a parameter when identifying the joint distribution of a set of random variables as negative multinomial.

We next provide some important properties of the NMD. The marginal distribution of each $Y_i, i = 1,2, \dots, h$, is negative binomial with parameters ν and $\frac{p_0}{p_0 + p_i}$, i.e. $Y_i \sim NB\left(\nu, \frac{p_0}{p_0 + p_i}\right)$. The proof (given in terms of the marginal distribution of Y_1) is found in Steyn (1959) and is as follows:

$$\sum_{y_2} \cdots \sum_{y_h} \frac{\left(\nu + y_1 + \sum_{i=2}^{h} y_i - 1\right)!}{(\nu-1)!\, y_1!\, y_2! \cdots y_h!} p_0^{\nu} p_1^{y_1} p_2^{y_2} \cdots p_h^{y_h}$$

$$= \frac{(\nu + y_1 - 1)!}{(\nu-1)!\, y_1!} p_1^{y_1} \sum_{y_2} \cdots \sum_{y_h} \frac{\left(\nu + y_1 + \sum_{i=2}^{h} y_i - 1\right)!}{(\nu + y_1 - 1)!\, y_2! \cdots y_h!} p_0^{\nu} p_2^{y_2} \cdots p_h^{y_h}$$

$$= \frac{(\nu + y_1 - 1)!}{(\nu - 1)!\, y_1!} p_1^{y_1} p_0^{\nu} \left(1 - \sum_{i=2}^{h} p_i\right)^{-(\nu + y_1)} = \frac{(\nu + y_1 - 1)!}{(\nu - 1)!\, y_1!} \left(\frac{p_0}{p_0 + p_1}\right)^{\nu} \left(\frac{p_1}{p_0 + p_1}\right)^{y_1}$$

$$= \binom{\nu + y_1 - 1}{y_1} \left(\frac{p_0}{p_0 + p_1}\right)^{\nu} \left(1 - \frac{p_0}{p_0 + p_1}\right)^{y_1} \Rightarrow Y_1 \sim NB\left(\nu, \frac{p_0}{p_0 + p_1}\right)$$

Similarly, the marginal distribution for the remaining $Y_i, i = 2, \ldots, h$, can be determined by replacing y_1 with y_i and adjusting the limits of the summations in the computations above to exclude either y_i or p_i as appropriate. These results imply that the marginal expected value of Y_i is $\nu \frac{p_i}{p_0}$ and the marginal variance of Y_i is $\nu \frac{p_i(p_0 + p_i)}{p_0^2}$. Furthermore, it was shown that the covariance of and correlation between Y_i and Y_j for $i \neq j$ is given by $Cov(Y_i, Y_j) = \nu \frac{p_i p_j}{p_0^2}$ and $Corr(Y_i, Y_j) = \sqrt{\frac{p_i p_j}{(p_0 + p_i)(p_0 + p_j)}}$ (Steyn et al., 1989)[1].

1.2: Brief History of the Negative Multinomial Distribution

According to Sibuya, Yoshimura, and Shimizu (1964), "the notion of the negative multinomial distribution was first introduced in the model of the inverse sampling[2] in multiple Bernoulli trials…", and the first systematic analysis of the NMD is attributed to Bates and Neyman (1952), who referred to the distribution as the multivariate negative binomial distribution[3] and derived its probability mass function via the probability generating function. While studying the theory of accident proneness, Bates and Neyman derived the NMD by considering s kinds of accidents, in which one type of accident is

[1] Note that the correlation between Y_i and Y_j is positive in a negative multinomial distribution, whereas the correlation between random variables in a multinomial distribution is negative.

[2] Casella and Berger (2002) describe inverse sampling techniques as sampling until r individuals with a certain characteristic are obtained from a population in which the proportion of individuals possessing the characteristic is p (pages 96-97).

[3] Though both the terms "negative multinomial" and "multivariate negative binomial" can be found in the literature, Johnson, Kotz, and Balakrishnan (1997, page 98) suggest that the term "negative multinomial" is a more accurately descriptive name for the distribution, and, as such, we will use this term exclusively throughout this dissertation following the historical overview in Section 1.2.

classified as a severe accident and the remaining $s-1$ are classified as different types of light predictor accidents. The random variables $X_1, X_2, \dots, X_s$ are used to represent the number of each type of aforementioned accident, and it is assumed that these random variables are mutually independent, each following a Poisson law with parameter $a_i\lambda, i = 1,2, \dots, s$, where to each individual of the population there is associated a positive number λ measuring his or her proneness to accidents. If λ follows a distribution Λ with density function $p_\Lambda(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, then the joint distribution of $X_1, X_2, \dots, X_s$ is negative multinomial (or as Bates and Neyman termed it, an s-variate negative binomial distribution) with probability mass function of the form

$$P\{(X_1 = n_1)(X_2 = n_2)\cdots(X_s = n_s)\} = \left[1 + \sum_{i=1}^{s} b_i\right]^{-\alpha} \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \prod_{i=1}^{s} \frac{c_i^{n_i}}{n_i!}$$

where $n = \sum_{i=1}^{s} n_i$, $b_i = \frac{a_i}{\beta}$, and $c_i = \frac{b_i}{1+\sum_{j=1}^{s} b_j} = \frac{a_i}{\beta+\sum_{j=1}^{s} a_j}$ for $i = 1,2, \dots, s$.

Bates and Neyman note that when this model is applicable, the s-dimensional problem can be reduced to a two-dimensional problem by letting X denote the number of severe accidents and Y denote the total number of light accidents (i.e. Y incorporates all $s-1$ original types of light accidents). The authors subsequently discuss estimation of the parameters in the resulting bivariate negative binomial distribution. Bates and Neyman also prove the following properties:

(i) The marginal joint distribution of a group of m variables, say $X_1, X_2, \dots, X_m$, is an m-variate negative binomial distribution

(ii) The joint distribution of $X_1, X_2, \dots, X_m$ and the sum $\chi = X_{m+1} + \cdots + X_s$ is an $(m+1)$-variate negative binomial distribution

(iii) The conditional joint distribution of $X_1, X_2, \ldots, X_m$ given values for the remaining $s - m$ variables is an m-variate negative binomial distribution and depends only on the value $\chi = x_{m+1} + \cdots + x_s$

Additional properties of the NMD have been given by many authors, including Sibuya, Yoshimura, and Shimizu (1964) and Nguyen et al. (2007), and a thorough treatment of the distribution can be found in Johnson, Kotz, and Balakrishnan (1997, pages 93-123).

1.3: Definition of Balanced and Unbalanced Negative Multinomial Distributions

In this dissertation, we will distinguish between "balanced" and "unbalanced" negative multinomial distributions. For an experiment in which there are $h + 1$ possible outcomes that can be modeled by an NMD, we define a balanced negative multinomial distribution as one in which the probability that each of the $h + 1$ outcomes occurs is equal (i.e. $p_i = \frac{1}{h+1}, i = 0,1, \ldots, h$). When the relationship $p_0 = p_1 = \cdots = p_h$ does not hold, the NMD will be referred to as an unbalanced negative multinomial distribution. A subset of the unbalanced distributions which may be of special interest is when $p_0 \neq p_1 = p_2 = \cdots = p_h$, and we term these "partially balanced" negative multinomial distributions. We next provide examples of balanced, unbalanced, and partially balanced negative multinomial distributions.

As an example of the balanced NMD, consider a fair six-sided die. If we roll the die until we observe ten 6's, then the distribution of the number of 1's, 2's, 3's, 4's and 5's observed during the rolls (denoted by $Y_1, \ldots, Y_5$, respectively) follows a balanced negative multinomial distribution with parameters 10 and $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = \frac{1}{6}$ and probability mass function

$$\left(10-1+\sum_{i=1}^{5} y_i\right)!\frac{\left(\frac{1}{6}\right)^{10}}{(10-1)!}\prod_{i=1}^{5}\frac{\left(\frac{1}{6}\right)^{y_i}}{y_i!}=\left(9+\sum_{i=1}^{5} y_i\right)!\frac{\left(\frac{1}{6}\right)^{10+\sum_{i=1}^{5} y_i}}{9!\prod_{i=1}^{5} y_i!}$$

Now, suppose that each time the die is rolled the probability that the i^{th} face is observed is proportional to the number on the face of the die (i.e. the probability that a 1, 2, 3, 4, 5, and 6 occurs is $\frac{1}{21},\frac{2}{21},\frac{3}{21},\frac{4}{21},\frac{5}{21}$, and $\frac{6}{21}$, respectively). Suppose we again roll the die until ten 6's are observed. In this case, the NMD is unbalanced with probability mass function

$$\left(10-1+\sum_{i=1}^{5} y_i\right)!\frac{\left(\frac{6}{21}\right)^{10}}{(10-1)!}\frac{\left(\frac{1}{21}\right)^{y_1}}{y_1!}\frac{\left(\frac{2}{21}\right)^{y_2}}{y_2!}\frac{\left(\frac{3}{21}\right)^{y_3}}{y_3!}\frac{\left(\frac{4}{21}\right)^{y_4}}{y_4!}\frac{\left(\frac{5}{21}\right)^{y_5}}{y_5!}$$

$$=\left(9+\sum_{i=1}^{5} y_i\right)!\frac{\left(\frac{6}{21}\right)^{10}}{9!}\prod_{i=1}^{5}\frac{\left(\frac{i}{21}\right)^{y_i}}{y_i!}$$

Finally, suppose a gambler carries a loaded die in which the probability a 6 is observed is $9/10$ and the probability a 1, 2, 3, 4, or 5 each occurs is $1/50$. If the gambler rolls the die until ten 6's are observed, then the NMD is partially balanced with mass function

$$\left(10-1+\sum_{i=1}^{5} y_i\right)!\frac{\left(\frac{9}{10}\right)^{10}}{(10-1)!}\prod_{i=1}^{5}\frac{\left(\frac{1}{50}\right)^{y_i}}{y_i!}=\left(9+\sum_{i=1}^{5} y_i\right)!\frac{\left(\frac{9}{10}\right)^{10}\left(\frac{1}{50}\right)^{\sum_{i=1}^{5} y_i}}{9!\prod_{i=1}^{5} y_i!}$$

1.4: Applications of the Negative Multinomial Distribution

Several applications of the negative multinomial distribution have been published in the literature since its introduction by Bates and Neyman. In this subsection, we will briefly discuss some of these examples[4] and provide references for additional applications.

[4] We omit many of the details required to derive the NMD in these examples as the purpose of Section 1.4 is solely to emphasize the usefulness of this distribution in real-world applications. Readers should consult the original cited articles for complete details.

Sinoquet and Bonhomme (1991) use the NMD to analyze radiation interception in a two-species plant canopy; in particular, they consider the interception of radiation coming from a given direction and going across a homogenous vegetation layer of thickness Z. Their approach consists of modeling the light relations that exist when two species of plants are planted in the same field, taking into account the geometrical structures in the vegetative canopies (i.e. the spatial distribution of the foliage elements of the two species). They define two components of leaf dispersion: within-species leaf dispersion (WSLD), which describes the rate of foliage overlap between leaves of plants of the same species, and between-species leaf dispersion (BSLD), which describes the rate of foliage overlap between leaves of different plant species. Leaf dispersion can be classified as regular (leaves avoid mutual shading), random, or clumped (leaves tend to overlap). Dividing the homogeneous layer Z into N equal and independent sublayers of thickness z (i.e. $Z = Nz$), Sinoquet and Bonhomme show that when the BSLD is regular, the probability of interception by species i ($i = 1,2$) is given by $p_i = 1 - e^{-k_i}$, where k_i is a function of the leaf area density and a projection coefficient onto a horizontal plane of a unit of leaf area of species i. The interception probabilities are described by a bivariate multinomial distribution with parameters N, p_1, and p_2. For the entire layer Z, the probability of no interception, P_0, is given by

$$P_0 = exp\left\{\frac{Z}{z}\ln[exp(-k_1 z) + exp(-k_2 z) - 1]\right\}$$

When the BSLD is clumped, the value of z in the expression above is taken to be negative, and the authors argue that this is justified by the use of a negative multinomial distribution to characterize the interception probabilities. A bivariate NMD is subsequently used to model the number of interceptions in the two plant species.

Patil and Boswell (1972) consider birth and death processes in which the corresponding rates factor into a function of time and a function of the size of the population components. If $\{\boldsymbol{X}(t): t \in \mathbb{R}^+\}$ denotes an r-dimensional pure birth process with birth rates $\lambda_k(\boldsymbol{x}, t)$, then the process is characterized by the differential equations

$$\frac{\partial}{\partial t} p(\boldsymbol{x}, t) = \sum_{k=1}^{r} \lambda_k(\boldsymbol{x} - i_k, r) p(\boldsymbol{x} - i_k, t) - \lambda(\boldsymbol{x}, t) p(\boldsymbol{x}, t)$$

where $p(\boldsymbol{x}, t) = P[\boldsymbol{X}(t) = \boldsymbol{x}], \lambda(\boldsymbol{x}, t) = \sum_{k=1}^{r} \lambda_k(\boldsymbol{x}, t)$ and i_k is a vector with a 1 in the k^{th} position and zero for all other positions. Patil and Boswell prove that if $\{\boldsymbol{X}(t): t \in \mathbb{R}^+\}$ is a birth process with $\boldsymbol{X}(0) = \boldsymbol{0}$, $\lambda_k(\boldsymbol{x}, t) = a_k(\gamma + \delta \boldsymbol{x} \cdot \boldsymbol{1}) h(t)$, $\sum_{k=1}^{r} a_k = 1$, and $\delta = 1$, then $\boldsymbol{X}(t)$ has the negative multinomial distribution with mass function

$$p(\boldsymbol{x}, t) = \binom{\gamma + \boldsymbol{x} \cdot \boldsymbol{1} - 1}{\boldsymbol{x}} p_0^{\gamma}(t) \prod_{k=1}^{r} \{a_k [1 - p_0(t)]\}^{x_k}$$

where $p_0(t) = e^{-\int_0^t h(s) ds}$.

Engel (1986) considers a model for count data in a split-plot design with two whole plot factors A and B (indexed by i and j, respectively) and one sub-plot factor C (indexed by k) with an equal number of replicates per cell (indexed by l). Assuming whole plot error, interaction between sub-plot factor C and whole plot error, sub-plot error, and a Poisson distribution as the basis of the model for X_{ijkl} (the count response for replicate l of sub-plot k of whole plot (i, j)), Engel posits the following model for X_{ijkl}:

(i) $X_{ijkl} \sim Poisson(m_{ijkl})$ with m_{ijkl} an observed value of the positive random variable M_{ijkl}

(ii) $M_{ijkl} = G_{ijl}(\alpha_{ij}, \theta) \cdot H_{ijkl}(\beta_{ijk})$ where the random variables G are independently Gamma distributed and $H_{ijkl}(\beta_{ijk}) = G_{ijkl}(\beta_{ijk}, \eta) / \sum_k G_{ijkl}(\beta_{ijk}, \eta)$

Omitting the subscripts i, j, and l leads to a negative multinomial distribution for the vector $(X_1, X_2, \dots, X_K)$ given $H_k = h_k$ with probability mass function

$$\binom{x_+ + \alpha - 1}{x_1, \dots, x_K, \alpha - 1}\left(\frac{1}{1 + \theta \sum_k h_k}\right)^{\alpha} \prod_k \left(\frac{\theta h_k}{1 + \theta \sum_k h_k}\right)^{x_k}$$

Our next example of the NMD is an original application to the theory of quality control. Suppose that a certain machine used in manufacturing is subject to breakdown due to the failure of any one of $h + 1$ components. The machine breaks down due to component i with (unknown) probability $p_i, i = 0,1, \dots, h$. If we want to estimate the probability of breakdown due to each component, we may observe numerous machines until we observe ν breakdowns due to component 0 and then count the number of breakdowns that have occurred due to the h remaining types of components, denoted by $Y_1, \dots, Y_h$. This experiment follows an NMD with parameters $\nu, p_0, p_1, \dots, p_h$. This model could be particularly important if the 0^{th} component is very expensive to repair or replace relative to the other h components, and hence we may only be willing to allow a certain number of breakdowns due to failure of the 0^{th} component before terminating the experiment and estimating the probability of breakdown due to each component.

Derivation of the NMD from an urn model and from an inverse sampling scheme can be found in Sibuya, Yoshimura, and Shimizu (1964). The use of the NMD in inverse sampling schemes may be of particular importance in ecological capture-recapture experiments. In such situations, estimates with better sampling properties are obtained since it is guaranteed that a predetermined number of tagged individuals will be recaptured. This contrasts the use of direct sampling schemes which may result in a low number of recaptures, indicating the need for additional sampling (Ord, Patil, and Taillie,

1979, page 177). Other practical examples, with references, are listed on pages 95-96 of Johnson, Kotz, and Balakrishnan (1997).

1.5: The Negative Multinomial as a Mixture Distribution

Karlis and Xekalaki (2005) define a mixture distribution as follows:

> A probability distribution is said to be a mixture distribution if its distribution function $F(\cdot)$ can be written in the form $F(\cdot) = \int_{\Theta} F(\cdot\,|\lambda) dG(\lambda)$, where $F(\cdot\,|\lambda)$ denotes the distribution function of the component densities considered to be indexed by a parameter λ with distribution function $G(\lambda), \lambda \in \Theta$. (page 35)

This definition can also be presented in terms of probability density functions as $f(x) = \int_{\Theta} f(x|\lambda) g_{\lambda}(\lambda) d\lambda$. In this representation, $g(\cdot)$ is referred to as the mixing density. When $X|\lambda \sim Poisson(\lambda)$ (i.e. $f(x|\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$), the random variable X is said to follow a mixed Poisson distribution (Karlis and Xekalaki, 2005).

Suppose $X|\lambda \sim Poisson(\lambda)$, where λ follows a gamma distribution[5] with parameters α and β (i.e. the mixing distribution $g_{\lambda}(\lambda)$ is a gamma distribution). Then,

$$P(X = x) = \int_0^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}} d\lambda = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \frac{1}{x!} \int_0^{\infty} \lambda^{x+\alpha-1} e^{-\lambda/\left(\frac{\beta}{\beta+1}\right)} d\lambda$$

$$= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \frac{1}{x!} \Gamma(x+\alpha) \left(\frac{\beta}{\beta+1}\right)^{x+\alpha} = \frac{(x+\alpha-1)!}{(\alpha-1)!\,x!} \left(\frac{1}{\beta+1}\right)^{\alpha} \left(\frac{\beta}{\beta+1}\right)^{x}$$

$$= \binom{\alpha+x-1}{x} \left(\frac{1}{\beta+1}\right)^{\alpha} \left(1 - \frac{1}{\beta+1}\right)^{x}$$

Thus, the marginal distribution of X is negative binomial with parameters α and $\frac{1}{\beta+1}$. This result is attributed to Greenwood and Yule (1920) and can be found in Johnson, Kotz, and Kemp (1992, page 204) or Neyman (1965) who utilizes the probability generating function to obtain the result.

---

[5] The probability density function of a $Gamma(\alpha, \beta)$ random variable Q is $f(q) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} q^{\alpha-1} e^{-q/\beta}$.

The definitions above define univariate mixed distributions. The definitions extend naturally to characterize multivariate mixed distributions, and here we will provide a specific case of multivariate mixed Poisson distributions as presented in Sibuya, Yoshimura, and Shimizu (1964). Suppose $X_1, X_2, .., X_r$ are independent Poisson random variables with parameters $m\lambda_i, i = 1, \dots, r$, where m is an observed value of a random variable M. The joint conditional distribution of $X_1, \dots, X_r$ is then

$$P(X_1 = x_1, \dots, X_r = x_r | M = m) = \prod_{i=1}^{r} e^{-m\lambda_i} \frac{(m\lambda_i)^{x_i}}{x_i!}$$

If the distribution of M, i.e. the mixing distribution, is taken to be a gamma distribution with parameters k and a, then the joint distribution of $X_1, \dots, X_r$ is negative multinomial as shown below:

$$P(X_1 = x_1, \dots, X_r = x_r) = \int_0^{\infty} \left\{ \prod_{i=1}^{r} e^{-m\lambda_i} \frac{(m\lambda_i)^{x_i}}{x_i!} \right\} \frac{1}{\Gamma(k)a^k} m^{k-1} e^{-m/a} dm$$

$$= \frac{1}{\Gamma(k)a^k} \prod_{i=1}^{r} \frac{\lambda_i^{x_i}}{x_i!} \int_0^{\infty} e^{-m\left(\sum_{i=1}^{r} \lambda_i + \frac{1}{a}\right)} m^{k+\sum_{i=1}^{r} x_i - 1} \, dm$$

$$= \frac{1}{\Gamma(k)a^k} \prod_{i=1}^{r} \frac{\lambda_i^{x_i}}{x_i!} \int_0^{\infty} e^{\frac{-m}{\left(\frac{a}{1+a\sum_{i=1}^{r} \lambda_i}\right)}} m^{k+\sum_{i=1}^{r} x_i - 1} \, dm$$

$$= \frac{\Gamma(k + \sum_{i=1}^{r} x_i) \left(\frac{a}{1 + a\sum_{i=1}^{r} \lambda_i}\right)^{k+\sum_{i=1}^{r} x_i}}{\Gamma(k)a^k} \prod_{i=1}^{r} \frac{\lambda_i^{x_i}}{x_i!}$$

$$= \Gamma\left(k + \sum_{i=1}^{r} x_i\right) \frac{\left(\frac{1}{1 + a\sum_{i=1}^{r} \lambda_i}\right)^{k}}{\Gamma(k)} \prod_{i=1}^{r} \frac{\left(\frac{a\lambda_i}{1 + a\sum_{i=1}^{r} \lambda_i}\right)^{x_i}}{x_i!}$$

That the product of independent Poisson variates mixed with a gamma distribution follows an NMD is attributed to Bates and Neyman (1952) and can also be found in

Papageorgiou (1983), Joshi (1975), Ord, Patil, and Taillie (1979, pages 167-168), and Johnson, Kotz, and Balakrishnan (1997, pages 94-95) and is stated without proof in Zhou and Lange (2010). An explicit derivation of this fact will also be provided in Section 2.4 where it will be obtained in the context of the comparative Poisson process.

The NMD can also be obtained via mixture of multiple Poisson variates and a multivariate gamma mixing distribution. This result, which is beyond the scope of this dissertation, can be found in Ferrari, Letac, and Tourneret (2004) or Chatelain, Lambert-Lacroix, and Tourneret (2009).

SECTION 2: RELATIONSHIP BETWEEN THE COMPARATIVE POISSON PROCESS AND THE NEGATIVE MULTINOMIAL DISTRIBUTION

2.1: The Comparative Poisson Process

Lehmann and Romano (2005) write,

> A problem arising in many different contexts is the comparison of two treatments or of one treatment with a control situation in which no treatment is applied. If the observations consist of the number of successes in a sequence of trials for each treatment, for example the number of cures of a certain disease, the problem becomes that of testing the equality of two binomial probabilities. If the basic distributions are Poisson, for example in a comparison of the radioactivity of two substances, one will be testing the equality of two Poisson distributions. (page 124)

This dissertation focuses on the latter setting, the comparative Poisson process; that is, the comparison of two (or more) populations in which the event count in each is independently Poisson distributed. Though we will be strictly concerned with the comparative Poisson process, the model can also be applied to the binomial setting when the number of trials is large and the probability of event occurrence is small (and hence the binomial distribution is closely approximated by the Poisson distribution). This is a well-known result and can be found, for example, on pages 66-67 and 93-94 in Casella and Berger (2002). Additionally, though Lehmann and Romano present the comparative Poisson process in terms of comparison of treatments to each other or a treatment to a control, and though this method is typically applied to clinical trials, the method can be applied to any comparison in which event counts follow or can be approximated by Poisson distributions. As such, in the remainder of Section 2, rather than use the terms "treatment" and "control" groups, we will use more general terminology to describe studies in which one or more "comparator" situations/groups are compared to one another or are compared to a "reference" situation/group. The specific application of comparative

Poisson methodology to clinical trials will be discussed at length in Sections 4 and 5 of this dissertation.

Results concerning the construction of uniformly most powerful unbiased (UMPU) tests for the parameters in exponential families, originally derived by Lehmann and Scheffé (1955), can be used to show that for $X \sim Poisson(\lambda)$ independent of $Y \sim Poisson(\mu)$, the UMPU test for the hypotheses $\mu \leq \lambda$ (or $\mu = \lambda$) versus $\mu > \lambda$ and for $\mu = \lambda$ versus $\mu \neq \lambda$ is based on the conditional distribution of Y given $T = X + Y$ (see also Lehmann and Romano (2005), pages 119-125). This conditional distribution is binomial as was first shown by Przyborowski and Wilenski (1940), and the derivation is reproduced below:

$$P(Y = y|X + Y = t) = \frac{P(Y = y, X = t - y)}{P(X + Y = t)} = \frac{\frac{e^{-\mu}\mu^{y}}{y!}\frac{e^{-\lambda}\lambda^{t-y}}{(t-y)!}}{\frac{e^{-(\mu+\lambda)}(\mu+\lambda)^{t}}{t!}}$$

$$= \frac{t!}{y!\,(t-y)!}\left(\frac{\mu}{\mu+\lambda}\right)^{y}\left(\frac{\lambda}{\mu+\lambda}\right)^{t-y} = \binom{t}{y}\left(\frac{\mu}{\mu+\lambda}\right)^{y}\left(1-\frac{\mu}{\mu+\lambda}\right)^{t-y}$$

We next provide examples of the application of the comparative Poisson process. Birnbaum (1953) considers continuous inspection of manufactured materials (such as cloth, paper, or wire) for flaws. The number of faults x observed over a length t of material may follow a Poisson distribution with parameter λt, where λ is the mean number of faults per unit length of material. Comparing the mean number of flaws per unit of two types of material is equivalent to comparing the parameters λ_1 and λ_2 of the two Poisson processes. Another example from Birnbaum (1953) is based on the use of a Geiger counter to observe the number of emissions x from a radioactive substance over time t. We assume that the distribution of events during the time interval is Poisson with

parameter λt, where λ represents the mean number of emissions per unit interval of time. If two radioactive substances are under study, then comparison of the parameters λ_1 and λ_2 is equivalent to comparison of the emission rates of the two substances.

Rather than compare two radioactive substances, we may want to evaluate the effectiveness of a shield designed to protect against radiation. To do so, we may introduce a steady source of radiation and record the number of emissions detected when the shield blocks the Geiger counter and again when the shield is removed. Comparing λ_2, the intensity of radiation when the shield is utilized, to λ_1, the intensity when the shield is removed, is statistically equivalent to the examples above when the number of emissions follows a Poisson distribution under both shielding conditions (i.e. presence or absence of the shield) (Birnbaum, 1954). In this example, the absence of the shield may be considered the "reference" situation and the presence of the shield the "comparator" situation.

Birnbaum's final example in the 1953 paper is to consider the number of cases of a rare disease observed among two large groups of individuals during a certain time period. If the number of cases of disease is independently Poisson distributed among the two populations, then the comparative Poisson model can be implemented to compare the incidence rates of disease in the two populations. An example of this application to disease incidence may be found in Hill, Spicer, and Weatherall (1968)[6] and is provided in Gail (1974). If the incidence of congenital malformations in a uranium mining town and in a control population follow Poisson distributions and i_2 and i_1 represent the incidence rates of malformations in the two populations, respectively, then the Poisson parameters

[6] See Gail (1974) for the reference to Hill, Spicer, and Weatherall (1968).

are $\lambda_2 = i_2 t$ and $\lambda_1 = i_1 t$, where t is the duration of observation. Testing the hypotheses $H_0: \lambda_2 = \lambda_1$ versus $H_a: \lambda_2 > \lambda_1$ is equivalent to testing $H_0: i_2 = i_1$ versus $H_a: i_2 > i_1$.

2.2: Designs for Comparing Two Poisson Populations

Gail (1974) proposes two designs, termed Design A and Design B, to conduct tests of two population parameters when the number of events in the two populations independently follow Poisson distributions with parameters λ_1 and λ_2, respectively. Specifically, Gail provides tests of $H_0: \lambda_2 = \lambda_1$ versus $H_a: \lambda_2 > \lambda_1$ (or equivalently $H_0: \rho = 1$ versus $H_a: \rho > 1$ where $\rho = \lambda_2/\lambda_1$)[7]. Design A is to observe the two populations, denoted by i ($i = 1, 2$), until a predetermined total number of events $T = X_1 + X_2$ has occurred, where X_i is the number of events observed in population i. Design B is to observe the two populations for a predetermined length of time, t[8]. The test under both designs is based on the conditional binomial distribution discussed in Section 2.1. The advantage of Design A is that an appropriate choice of T will always yield a critical region of sufficient power, though the disadvantage is that a study termination date cannot be specified. Thus, it could take a significant amount of time for a study under Design A to terminate, which also implies significant expenses. Design B does have a specified termination date at time t, but if few events have occurred among the populations at this time, a critical region of insufficient power may result.

Some alternative tests for comparing two Poisson parameters can be found in Birnbaum (1954). Birnbaum provides a test in terms of $\gamma = \lambda_1/\lambda_2$ based on the F

[7] Gail (1974) provides computations for Designs A and B under the assumption of equal population sizes. The designs were extended by Brown and Green (1982) to the case of unequal population sizes.

[8] Note that the concept of time in Gail's paper may be readily substituted to appropriately reflect the comparative Poisson process under study. For example, we previously considered continuous inspection of manufactured materials, and in this setting, t would represent the length of material examined. In the case of a rare binomial event that is approximated by the Poisson distribution, such as a vaccine study with a rare adverse event as the primary outcome, t would represent the number of subjects in the study.

distribution, which is particularly appealing when the Poisson processes are separated in space or time, and a test in terms of $\Delta = \lambda_1 - \lambda_2$ based on ranking two exponential populations with respect to their means (see the original article for details and for additional testing approaches).

Here we consider "Design C"[9] in which the study is terminated once d_C events have been observed in population 1 (the "reference" population) and we record the number of events that have occurred in population 2 (the "comparator" population) by the time of stoppage. Since the waiting time for a single event in a Poisson process follows an Exponential distribution, the waiting time for d_C events to occur is the sum of independent, identically distributed Exponential variables, which follows a Gamma distribution (see for example Gallager (1996), pages 33-36, for a formal proof). Hence, tests of the event incidence rates (or equivalently, the Poisson parameters) in the two populations can be derived from a mixture of a Poisson and Gamma distribution, which we have shown in Section 1.5 is a negative binomial distribution[10].

Like Design A, a study conducted under Design C will always result in a critical region of sufficient power (given that an appropriate value of d_C is chosen), but the duration of the study cannot be specified[11].

---

[9] To our knowledge, Design C was first proposed by Hsu (2010, pages 86-87).

[10] In Sections 4 and 5, we will show that when multiple comparator groups are compared to a reference group and the number of events that occur in each group independently follows a Poisson distribution conditional on the time elapsed in the reference group, a test of equivalency of the Poisson parameters between the comparator and reference groups can be conducted based on the negative multinomial distribution. This is based on the fact that the mixture of multiple independent Poisson variates with a gamma distribution follows an NMD, which was shown in Section 1.5. When only one comparator group is under study (i.e. one Poisson variate mixed with a gamma distribution), the NMD reduces to the negative binomial distribution as was shown in Section 1.1.

[11] Note that Design C is equivalent to Design A when the following two conditions are satisfied: (1) there are a total of two groups under study (for example, one comparator group and one reference group) and (2) when curtailment is applied to the study (see Section 4.5 for a description of curtailment).

2.3: Comparison of Multiple Poisson Populations

Thus far, we have discussed comparative Poisson designs in which only two populations are under study. Naturally, an extension of the comparative Poisson process to greater than two populations is of interest. Examples of studies comparing more than two populations in which events accrue according to Poisson processes are provided below.

Consider again the example of determining the effectiveness of a shield in protecting against radiation emitted by a substance. Suppose now that there are K shields made from different types of material, and we are interested in comparing the effectiveness of these shields in relation to the reference situation (i.e. absence of a shield in front of the Geiger counter). Here, we are interested in comparing multiple shields (i.e. multiple comparators) to a single reference situation.

Peng and Krishnamoorthy (2010) collect and present several examples of comparative Poisson processes with greater than two populations; they are as follows: Nelson, Wludyka, and Copeland (2005) suggests an example in which the arrival rates of patients to six urgent clinics run by a health maintenance organization are compared using samples of arrival counts from each clinic. Brown and Zhao (2002) describe a situation in which the average number of service request calls per day is compared among several call centers. Chiu and Wang (2009) consider comparison of the death rates of patients in four groups following heart valve replacement[12].

Two final examples come from Singh (1980), who suggests that

> …an air pollution research study might involve exposing the plant Tradescantia to several levels of polluted air samples and comparing counts of mutants from the polluted air samples with mutant counts from a control sample. In another application,

[12] See Peng and Krishnamoorthy (2010) for the references to Nelson, Wludyka, and Copeland (2005), Brown and Zhao (2002), and Chiu and Wang (2009).

> a scientist may be interested in comparing the counts of surviving bacteria colonies in treated groups at several levels with those from the control. (page 1138)

These examples indicate the need for comparative Poisson designs for several comparator groups (perhaps in relation to a single reference group). Researchers have proposed a variety of tests appropriate for this situation. Firstly, we consider the extension of Gail's Designs A and B for testing directional hypotheses. It should be unsurprising that the multivariate test is based on the multivariate extension of the binomial distribution, that is, the multinomial distribution. Hsu (2010) provides a thorough treatment of these tests in the context of clinical trials which compare multiple treatment groups to a single control group. Alternatively, Suissa and Salmi (1989) provide test statistics based on unidirectional Z statistics for comparing several exposed groups to a single reference group and for comparing one exposed group to several reference groups. Finally, Singh (1980) implements a Bayesian framework to evaluate $H_0: \lambda_0 = \lambda_1 = \cdots = \lambda_k$ versus $H_0: \lambda_0 \neq \lambda_j$ for at least one value of j, where λ_0 is the Poisson parameter of a control group and $\lambda_1, \ldots, \lambda_k$ are the Poisson parameters of k treatment groups.

If we wish to test the non-directional hypotheses $H_0: \lambda_1 = \cdots = \lambda_m$ versus $H_a: \lambda_i \neq \lambda_j$ for some $i \neq j$, then it can be shown that for $Y_i \sim Poisson(n_i\lambda_i)$, the distribution of $Y_1, \ldots, Y_m | \sum_{i=1}^{m} Y_i = T$ is multinomial with probability mass function $\frac{T!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m}$, where $p_i = \frac{n_i\lambda_i}{\sum_{j=1}^{m} n_j\lambda_j}$ (Peng and Krishnamoorthy, 2010). Thus, the exact conditional test of H_0 versus H_a can be conducted by calculating multinomial probabilities, and the test is UMPU (Suissa and Salmi, 1989). However, the most common test for comparing the underlying event rates among several Poisson

populations is based on the chi-squared distribution. Again using the notation of Peng and Krishnamoorthy, the test statistic $\chi^2 = \sum_{i=1}^{m} \frac{n_i(\hat{\lambda}_i - \hat{\lambda})^2}{\hat{\lambda}}$ follows a chi-squared distribution with $m - 1$ degrees of freedom as $n_i \to \infty$, where $\hat{\lambda}_i = \frac{Y_i}{n_i}$ and $\hat{\lambda} = \frac{\sum_{i=1}^{m} Y_i}{\sum_{i=1}^{m} n_i}$. Peng and Krishnamoorthy also propose a parametric bootstrap test and compare its results to both the exact conditional test based on the multinomial distribution and the approximate test based on the chi-squared distribution. Brown and Zhao (2002) have proposed a test based on Anscombe's variance stabilizing transformation, though it is not applicable to the case of unequal sample sizes.

Section 2.4: Multivariate Extension of Design C

As described in Section 2.3, when we considered the multivariate extension of Designs A and B we obtained a test based on the multinomial distribution, the multivariate counterpart to the binomial distribution. Applying the same reasoning, we should expect that the multivariate extension of Design C should result in a test based on the negative multinomial distribution. Here we will explicitly derive the NMD from a comparative Poisson framework, as was alluded to in Section 1.5. To do so, we require the notation provided in Table 1 below and the preliminary results from Hsu (2010, pages 86-87) which follow.

Table 1: Notation

K	Number of comparator groups (does not include the reference group)
d_C	Number of events to observe in the reference group
D_k	Number of events in the k^{th} comparator group, $k = 1,2, \dots, K$
N_C, t	Number of person years in the reference group to reach d_C events; t is an observed value of N_C
N_{T_k}	Number of person years in the k^{th} comparator group when the study arm or entire study terminates, $k = 1,2, \dots, K$
i_C	Incidence rate of events per person year in the reference group
i_k	Incidence rate of events per person year in the k^{th} comparator group, $k = 1,2, \dots, K$
$r_k = i_k/i_C$	Rate ratio of comparator group k to the reference group, $k = 1,2, \dots, K$
$\lambda_C = i_C N_C$	Poisson intensity rate in the reference group for a given N_C
$\lambda_k = i_k N_C$	Poisson intensity rate in the k^{th} comparator group for a given N_C

Note: N_{T_k} and r_k will be introduced in Section 4. All other notation in Table 1 is introduced in Section 2.4.

Suppose there are K comparator groups which we will compare to a single reference group. The incidence rate of events per person year[13] for the k^{th} comparator group is $i_k, k = 1,2, \dots, K$, and the incidence rate of events per person year in the reference group is i_C. We will terminate the study once the number of events observed in the reference group reaches a pre-specified number d_C and record the number of events that have occurred in each of the K comparator groups by the time of study stoppage, denoted by $D_1, D_2, \dots, D_K$. Event accrual in the reference group follows a Poisson distribution with parameter $\lambda_C = i_C N_C$, where N_C denotes the number of person years it takes the reference group to reach d_C events. Since we do not know many person years it will take the reference group to accrue d_C events (i.e. N_C is a random variable), the duration of the study is unknown. However, we know that the distribution of N_C is Gamma with parameters d_C and $1/i_C$, conditional on the value of d_C (i.e.

[13] Person years of follow-up is defined as the total amount of study-time contributed by all study participants. Here we have substituted the notion of "time to event occurrence" with that of "person years until event occurrence", as person years represents a more natural measure of duration in the context of clinical trials, which will be our primary focus in Sections 4 and 5.

$N_C|d_C \sim Gamma(d_C, 1/i_C)$). This can be understood by considering the number of person years until one event occurs in the reference group to have an Exponential distribution with parameter $1/i_C$ and applying the argument in Section 2.2 concerning the sum of Exponential random variables; a formal proof can be found in Appendix A[14]. Once the value of N_C is known, the number of events that occur in each of the K comparator groups follows a Poisson distribution with parameter $\lambda_k = i_k N_C, k = 1, \dots, K$, and the distribution of D_k no longer depends on d_C for $k = 1, \dots, K$ (i.e. D_k only depends on the number of person years needed to obtain d_C events in the reference group); as such, $D_k|N_C \sim Poisson(i_k N_C)$ for $k = 1, \dots, K$. So, we have $P(D_k = d_k|d_C, N_C = t) = P(D_k = d_k|N_C = t)$. Furthermore, conditional on $N_C = t$, the D_k are independent of one another. Hsu proves that the distribution of $D_1, \dots, D_K$ conditional on d_C is negative multinomial; the proof is reproduced below:

$$P(D_1 = d_1, \dots, D_K = d_k|d_C) = \int_0^\infty P(D_1 = d_1, \dots, D_K = d_k, N_C = t|d_C)\, dt$$

$$= \int_0^\infty P(D_1 = d_1, \dots, D_K = d_k|d_C, N_C = t) P(N_C = t|d_C)\, dt$$

$$= \int_0^\infty P(D_1 = d_1, \dots, D_K = d_K|N_C = t) P(N_C = t|d_C) dt$$

$$= \int_0^\infty \left[\prod_{k=1}^{K} P(D_k = d_k|N_C = t)\right] P(N_C = t|d_C) dt$$

[14] Note that the proof in Appendix A considers time intervals for event occurrence, but the notion of time in the proof may be readily substituted by that of person years in accordance with the terminology used throughout Section 2.4 and beyond in this dissertation.

$$= \int_0^\infty \left[\prod_{k=1}^{K} \frac{e^{-i_k t}(i_k t)^{d_k}}{d_k!} \right] \frac{i_C^{d_C}}{\Gamma(d_C)} t^{d_C - 1} e^{-i_C t} dt$$

$$= \int_0^\infty \left[\prod_{k=1}^{K} \frac{i_k^{d_k}}{d_k!} \right] e^{-t(\sum_{k=1}^{K} i_k)} t^{\sum_{k=1}^{K} d_k} \frac{i_C^{d_C}}{\Gamma(d_C)} t^{d_C - 1} e^{-i_C t} dt$$

$$= \int_0^\infty \left[\prod_{k=1}^{K} \frac{i_k^{d_k}}{d_k!} \right] t^{d_C + \sum_{k=1}^{K} d_k - 1} e^{-(i_C + \sum_{k=1}^{K} i_k)t} \frac{i_C^{d_C}}{\Gamma(d_C)} dt$$

$$= \frac{i_C^{d_C}}{\Gamma(d_C)} \left[\prod_{k=1}^{K} \frac{i_k^{d_k}}{d_k!} \right] \int_0^\infty t^{d_C + \sum_{k=1}^{K} d_k - 1} e^{-(i_C + \sum_{k=1}^{K} i_k)t} dt$$

$$= \frac{i_C^{d_C}}{\Gamma(d_C)} \left[\prod_{k=1}^{K} \frac{i_k^{d_k}}{d_k!} \right] \frac{\Gamma(d_C + \sum_{k=1}^{K} d_k)}{(i_C + \sum_{k=1}^{K} i_k)^{d_C + \sum_{k=1}^{K} d_k}}$$

$$= \Gamma\left(d_C + \sum_{k=1}^{K} d_k \right) \frac{\left(\frac{i_C}{i_C + \sum_{k=1}^{K} i_k} \right)^{d_C}}{\Gamma(d_C)} \prod_{k=1}^{K} \frac{\left(\frac{i_k}{i_C + \sum_{k=1}^{K} i_k} \right)^{d_k}}{d_k!}$$

Thus, by Equation 1.1.2, we have

$$D_1, D_2, \ldots, D_K | d_C \sim NM\left(d_C, \frac{i_C}{i_C + \sum_{k=1}^{K} i_k}, \frac{i_1}{i_C + \sum_{k=1}^{K} i_k}, \frac{i_2}{i_C + \sum_{k=1}^{K} i_k}, \ldots, \frac{i_K}{i_C + \sum_{k=1}^{K} i_k} \right) \quad (2.4.1)$$

In Section 2.2, we alluded to the fact that Design C may be preferred to Design B, as an appropriate choice of d_C will always yield a critical region of sufficient power. There are also two primary reasons why Design C may be preferred to Design A. Firstly, the independence of the D_k achieved by conditioning on N_C greatly simplifies the necessary calculations for establishing testing procedures (see Sections 4 and 5); this contrasts the lack of independence under the multinomial testing paradigm when Design A is applied to studies of multiple populations. Secondly, when multiple comparator groups and a single reference group are evaluated under Design A, it is possible that one

of the comparator groups will be responsible for the majority of the total number of events observed (this may happen, for example, if the incidence rate of events in this comparator group is underestimated during study planning). Such a situation will limit the amount of information available for drawing conclusions about the remaining comparator groups in relation to the reference group. This limitation will not apply to the testing procedures under Design C which we propose in subsequent sections of this dissertation.

In Sections 4 and 5, we will use the Design C framework and Equation 2.4.1 to design clinical trials in which multiple experimental treatments are compared to a single control treatment. The experimental treatment groups serve as the comparator groups discussed here in Section 2, and the control group similarly equates to the reference group. We will use the minimum and maximum of the D_k to compare the experimental treatments to the control; as such, we next discuss the order statistics of the NMD in Section 3.

SECTION 3: ORDER STATISTICS OF THE NEGATIVE MULTINOMIAL DISTRIBUTION

3.1: Definition of the Order Statistics of the Negative Multinomial Distribution

Consider a negative multinomial experiment in which there are $K+1$ possible outcomes denoted by $E_0, E_1, \dots, E_K$ which occur with probabilities $p_0, p_1, \dots, p_K$, respectively, and we conduct independent trials until outcome E_0 (which we refer to as the "reference" outcome to distinguish it from the K remaining "comparator" outcomes) occurs d_C times. Let X_k denote the number of trials that have resulted in outcome $E_k, k = 1,2, \dots, K$, by the time the process terminates at d_C occurrences of the reference outcome. Each X_k takes a value in $0,1,2, \dots$. If we arrange the X_k in ascending order and relabel the ordered variables as $X_{(1)}, X_{(2)}, \dots, X_{(K)}$, then we have defined the order statistics of the negative multinomial distribution[15]. For example, suppose we roll a fair die until we obtain five 6's (i.e. $d_C = 5$), and during the course of these trials we observe eight 1's, four 2's, five 3's, ten 4's, and seven 5's. Then our order statistics would be $X_{(1)} = 4, X_{(2)} = 5, X_{(3)} = 7, X_{(4)} = 8$, and $X_{(5)} = 10$. In the next two subsections, we will provide formulas to determine the distribution of the order statistics for a balanced negative multinomial distribution and consider the challenges in providing similar expressions for unbalanced negative multinomial distributions.

3.2: Order Statistics of a Balanced Negative Multinomial Distribution

To derive the distribution of the order statistics in a balanced NMD, we use the following theorem from Casella and Berger (2002, pages 227-228) concerning the order statistics of

[15] Notice that in our definition of the order statistics of the NMD we have excluded the fixed number of trials d_C for which the reference outcome E_0 is observed from consideration, and hence the number of order statistics in a $K+1$ outcome NMD is K. In Appendix B, we will extend the notion of the order statistics of the NMD to include the fixed value d_C, and we will therefore consider a $K+1$ outcome NMD to have $K+1$ order statistics.

discrete distributions:

Theorem 1: Let $X_1, \dots, X_n$ be a random sample from a discrete distribution with $P(X = x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible values of X in ascending order. Define

$$P_0 = 0$$
$$P_1 = p_1$$
$$P_2 = p_1 + p_2$$
$$\vdots$$
$$P_i = p_1 + p_2 + \cdots + p_i = P(X \leq x_i)$$
$$\vdots$$

Let $X_{(1)}, \dots, X_{(n)}$ denote the order statistics from the sample. Then

$$P\left(X_{(j)} \leq x_i\right) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

and

$$P\left(X_{(j)} = x_i\right) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right]$$

We can apply this theorem in conjunction with the comparative Poisson formulation of the NMD presented in Section 2.4 to find the distribution of the order statistics of balanced negative multinomial distributions (i.e. when the parameters $p_0 = p_1 = p_2 = \cdots = p_K = 1/(K + 1)$ in a $K + 1$ outcome NMD). The distribution must be balanced so that the random variables $X_1, \dots, X_K$ denoting the number of trials that result in each of the K comparator outcomes are identically distributed in accordance with the random sample requirement in Theorem 1[16].

Suppose we wait to observe d_C trials which result in the reference outcome in a $K + 1$ outcome NMD. Since $X_k, k = 1, \dots, K$, takes a value in the set 0,1,2, ..., per

[16] Independence of $X_1, \dots, X_K$ follows from the fact that conditional on $N_C = t$, the X_k are independent of one another, a fact that will be utilized in subsequent calculations; this property was discussed in Section 2.4.

Theorem 1 we have $x_1 = 0, x_2 = 1, \dots, x_i = i - 1, \dots$. From the comparative Poisson formulation of the NMD, we know that $P(X_k \le x_k|t) = \sum_{s=0}^{x_k} e^{-i_C t} \frac{(i_C t)^s}{s!}$ for $k = 1,2, \dots, K$ (as conditional on $N_C = t$, the number of trials resulting in each of the K comparator outcomes independently follows a Poisson distribution with parameter $\lambda_k = i_k t = i_C t$ in a balanced distribution). As $P_i = P(X \le x_i)$ for $i = 1,2, \dots$ in Theorem 1, we thus have $P_i = \sum_{s=0}^{x_i = i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}$. Hence, for the balanced negative multinomial distribution we may write

$$P\left(X_{(j)} \le x_i|t\right) = \sum_{l=j}^{K} \binom{K}{l} P_i^l (1 - P_i)^{K-l}$$

$$= \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^{K-l}$$

$$\Rightarrow P\left(X_{(j)} \le x_i\right) =$$

$$\int_0^\infty \frac{(i_C t)^{d_C} t^{-1} e^{-i_C t}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^{K-l} dt$$

where we have applied the fact that $t \sim Gamma(d_C, 1/i_C)$. We will now show that this integral is invariant with respect to the value of i_C, i.e. we will prove

$$\int_0^\infty \frac{(i_C t)^{d_C} t^{-1} e^{-i_C t}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^{K-l} dt$$

$$= \int_0^\infty \frac{x^{d_C - 1} e^{-x}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-x} \frac{x^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-x} \frac{x^s}{s!}\right]^{K-l} dx$$

To prove the equality, we make the following change of variables:

$$x = i_C t \Rightarrow t = \frac{x}{i_C}$$
$$dx = i_C dt \Rightarrow dt = \frac{dx}{i_C} \tag{3.2.1}$$

Then,

$$\int_0^\infty \frac{(i_C t)^{d_C} t^{-1} e^{-i_C t}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^{K-l} dt$$

$$= \int_0^\infty \frac{x^{d_C} \left(\frac{i_C}{x}\right) e^{-x}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-x} \frac{x^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-x} \frac{x^s}{s!}\right]^{K-l} \frac{dx}{i_C}$$

$$\Rightarrow P\left(X_{(j)} \leq x_i = i - 1\right)$$
$$= \int_0^\infty \frac{x^{d_C - 1} e^{-x}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i-1} e^{-x} \frac{x^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-x} \frac{x^s}{s!}\right]^{K-l} dx \tag{3.2.2}$$

Thus, the integral is invariant to the value of i_C. The formula for $P\left(X_{(j)} = x_i\right)$ follows directly as in Theorem 1 (i.e. by writing $P\left(X_{(j)} \leq x_i\right) - P(X_{(j)} \leq x_{i-1})$ and simplifying the resulting expression).

We have written functions in R to compute $P\left(X_{(j)} \leq i\right)$ and $P\left(X_{(j)} = i\right)$ for the balanced NMD. The function *balanced_order_less* takes the arguments dc (number of trials resulting in the reference outcome to be observed), j (denotes the j[th] order statistic), i (takes a value in 0,1,2, ...), and K (number of comparator outcomes in the experiment, i.e. not including the reference outcome). *balanced_order_less* returns $P(X_{(j)} \leq i)$. The function *balanced_order_equal* takes the same arguments as *balanced_order_less* and returns $P(X_{(j)} = i)$. For example, suppose we have a balanced distribution with $K = 5$ comparator outcomes and we wait to observe $d_C = 10$ reference outcomes (i.e. $K + 1 =$

6 total possible outcomes). Suppose we wish to determine the probability that the fourth order statistic (i.e. $j = 4$) is less than or equal to 4. *balanced_order_less(10,4,4,5)* tells us that this probability is 0.01403157. The probability that $X_{(4)} = 4$ is 0.01031401 according to *balanced_order_equal(10,4,4,5)*. Full code for *balanced_order_less* and *balanced_order_equal* can be found in Appendix E.

3.3: Order Statistics of an Unbalanced Negative Multinomial Distribution

When the NMD is unbalanced, we cannot apply Theorem 1 in a simple fashion to derive formulas for the order statistics of the distribution due to the lack of identical variables. However, it is simple to use simulation to calculate the desired probabilities. The R function *unbalanced_order* takes the arguments *probs* (vector of length K, where K is the number of comparator outcomes in the NMD, containing the probabilities of a trial resulting in each comparator outcome, i.e. not including the reference outcome), dc (number of trials resulting in the reference outcome to be observed), j (denotes the jth order statistic), i (takes a value in 0,1,2, ...), and *sims* (number of simulations used to estimate the probability). *unbalanced_order* is based on the R package 'MGLM' written by Zhang and Zhou (2017) and returns an estimate of $P(X_{(j)} \leq i)$ for an unbalanced negative multinomial distribution based on a user-selected number of simulations. For example, if we want to find the probability that the third order statistic is less than or equal to 4 when there are five comparator outcomes with underlying probabilities 0.1, 0.1, 0.3, 0.2, 0.1 (so the probability the reference outcome is observed in a trial is 0.2), and we conduct trials until we observe 10 reference outcomes, then *unbalanced_order(c(.1,.1,.3,.2,.1),10,3,4,1000000)* returns a probability of 0.218617 based on 1,000,000 simulations. To compute $P(X_{(j)} = i)$, simply compute $P\big(X_{(j)} \leq i\big) -$

$P\left(X_{(j)} \leq i-1\right)$ via two applications of *unbalanced_order*. Full code for *unbalanced_order* can be found in Appendix E.

SECTION 4: APPLICATION OF THE NEGATIVE MULTINOMIAL DISTRIBUTION TO COMPARATIVE POISSON SUPERIORITY CLINICAL TRIALS OF MULTIPLE EXPERIMENTAL TREATMENTS VERSUS A SINGLE CONTROL TREATMENT

4.1: Objectives

Section 2 provided several examples in which comparison of multiple Poisson rates, perhaps in relation to that of a single reference group, was of interest. Hsu (2010) proved that under Design C, testing of the rates can be based on a negative multinomial distribution. In particular, if we let $D_1, \dots, D_K$ represent the number of events observed in K comparator groups, and we wait until d_C events have been observed in a chosen reference group to terminate the study, then the conditional distribution $D_1, D_2, \dots, D_K | d_C$ is negative multinomial with parameters $d_C, \frac{i_C}{i_C+\sum_{k=1}^{K} i_k}, \frac{i_1}{i_C+\sum_{k=1}^{K} i_k}, \frac{i_2}{i_C+\sum_{k=1}^{K} i_k}, \dots, \frac{i_K}{i_C+\sum_{k=1}^{K} i_k}$ when the event accrual in each comparator population is conditionally Poisson distributed with parameter $i_k N_C$ and N_C, the number of person years to acquire d_C events in the reference group, follows a Gamma distribution with parameters d_C and $1/i_C$. In this section, we will consider the application of this result to clinical trials in which K experimental or new[17] treatment groups ($K \geq 1$) are compared to a single control treatment group. Hence, $D_1, \dots, D_K$ and $i_1, \dots, i_K$ will now represent the number of events (for example, the number of cases of disease) and the event incidence rates per person year in the K experimental treatment groups, and d_C and i_C will represent the corresponding values for the control group. The investigation of several treatments typically occurs early in the drug process (i.e. during drug discovery or phase 1 trials) when several similar molecular compounds or varying doses of one experimental agent

[17] We will use the terms "experimental treatment" and "new treatment" interchangeably throughout this dissertation.

are compared for efficacy and/or safety; however, it may also occur during late-phase trials when, for example, several new approved treatments are compared to either no treatment or to a standard of care treatment for efficacy and/or safety.

The structure of this section is as follows: Section 4.2 presents global hypotheses for testing the superiority of multiple experimental treatments in comparison to a single control treatment using Design C methodology, provides definitions of Type I error and power based on these hypotheses, and proposes an appropriate test procedure. Section 4.3 illustrates the implementation of this test via application to a real-world study of influenza vaccines. Section 4.4 compares the design parameters obtained under the exact Design C approach to those resulting from using the Bonferroni method for multiple comparisons. Section 4.5 discusses the differences between an uncurtailed and curtailed trial in the context of Design C, and Section 4.6 concludes with estimation of the expected number of person years of follow-up until trial termination in an uncurtailed and curtailed trial. Many of these results will make use of the comparative Poisson formulation of the NMD and the minimum and maximum order statistics of the NMD, which were presented in Sections 2 and 3 of this dissertation, respectively.

4.2: Global Hypotheses, Test Statistic, and Definitions of Power

In large-scale controlled clinical trials there are usually two objectives:

1. Efficacy: Prove the experimental treatment(s) is superior in efficacy to the control treatment
2. Safety: Prove the experimental treatment(s) is safe for consumption (i.e. does not cause too many adverse events in comparison to the control treatment)

This section focuses on clinical trials in which the primary objective is proving the superior efficacy of one or more experimental treatments compared to a single control that may be either a placebo or a current standard of care treatment; testing for safety of experimental treatments will be addressed in Section 5. Our results will apply very naturally to the study of vaccines where the outcome of interest is the occurrence of disease or an adverse reaction and a large number of study participants are observed. Vaccines are traditionally referred to as interventions as opposed to treatments, but in this dissertation the term "treatment" will generally refer to any agent which improves the medical outcome under study, regardless of whether it is of pharmaceutical, biologic, or non-chemical origin. We will restrict our attention throughout to the case of equal allocation of study participants to the K experimental treatment arms and control arm. That is, the allocation ratio of $Tx_1 : Tx_2 : \ldots : Tx_K : Control$ will be $1 : 1 : \ldots : 1 : 1$.

Our objective will be to determine whether sufficient statistical evidence exists that at least one of the experimental treatments is superior to the control treatment. In a comparative Poisson trial designed to demonstrate superiority, the outcome observed is the number of events (e.g. cases of disease) that occur in each group under study. Hence, an experimental treatment will be found superior to the control treatment if the number of events that occur in the experimental treatment group is significantly less than the number of events observed in the control group. Based on these considerations, our global null and alternative hypotheses are

$$H_0 : i_1 = i_2 = \cdots = i_K = i_C \; versus \; H_a : i_1 = r_1 i_C, i_2 = r_2 i_C, \ldots, i_K = r_K i_C \quad (4.2.1)$$
where all $r_k \leq 1$ and at least one of the r_k is strictly less than 1

The values $r_1, \ldots, r_K$ will be referred to as the "rate ratios" of the experimental treatment groups to the control group and represent the amount by which each of the K experimental treatments reduces the event incidence relative to the control treatment.

We next present definitions of Type I error and power for our test of the hypotheses in Equation 4.2.1. A Type I error (rejecting the null hypothesis when the null hypothesis is true) occurs when we declare at least one experimental treatment to be statistically superior to the control, when in fact none of the experimental treatments are superior to the control. To illustrate, consider a study in which three experimental treatments, denoted by A, B, and C, are compared to a single control treatment. If none of A, B, and C are superior to the control, then the probability that at least one experimental treatment is falsely declared superior to the control is the probability that A and/or B and/or C is declared superior to the control. The global Type I error incurred for the hypotheses in Equation 4.2.1 will be termed "overall Type I error" and will be denoted by α_{ovr}[18].

A Type II error (accepting the null hypothesis when the null hypothesis is false) corresponds to failing to declare any experimental treatment superior to the control, when

[18] Note that if one or more of the experimental treatments are truly superior to the control, then it is not possible to commit a global Type I error as defined above. However, consider again the example of treatments A, B, and C and suppose that A and B are truly superior to the control while C is not. If we reject the null hypothesis but falsely conclude that only treatment C is superior and thus is solely responsible for the rejection, then though a Type I error is not made (since A and B are truly superior to the control and thus the global null hypothesis should be rejected), we have incorrectly declared treatment C to be superior, and hence have made a Type I error if treatment C is considered on an individual basis (i.e. externally of the global testing framework). We have also made individual Type II errors on treatments A and B as we incorrectly failed to find them superior. This example illustrates that in testing the global hypotheses in Equation 4.2.1, both Type I and Type II errors can be made when the experimental treatments are considered individually. It may be of interest to define a Type I error when only a proper subset, say $\delta < K$, of the experimental treatments are not superior to the control. We call this "sub-Type I error" and define it to be the probability that at least one of the δ truly non-superior experimental treatments are incorrectly declared superior to the control. It is clear that the probability of a sub-Type I error is less than the probability of an overall Type I error as there are fewer non-superior treatments which can be incorrectly declared superior to the control.

in fact one or more experimental treatments are statistically superior to the control. Power is the probability that the null hypothesis is rejected when it is in fact false, and hence power is equal to $1 - P(Type\ II\ error)$. We will let β_{T_k} represent making a Type II error on the k^{th} individual experimental treatment, $k = 1,2, \dots, K$[19]. We define "pointwise power" to be the probability that an experimental treatment which has a given rate ratio r will be found to be superior to the control. Again, consider a study in which three experimental treatments, A, B, and C, are compared to a single control. Pointwise power answers the following question: if A is truly superior to the control with a rate ratio of r under the alternative hypothesis in Equation 4.2.1, then what is the probability that it will be found to be superior (and similarly for treatments B and C)? Pointwise power is related to the individual Type II errors via

$$Pointwise\ Power = 1 - P\left(\beta_{T_k}\right)$$

To develop a test for the hypotheses in Equation 4.2.1 using the Design C framework, we need to consider the practical implications of the definitions of Type I and Type II error given above. In clinical trials where the primary objective is to establish superiority of one or more new treatments, the permissible Type I error is usually fixed by a regulatory agency, as a Type I error corresponds to consumers' risk (since it corresponds to the probability that one or more experimental treatments are incorrectly declared superior to the control), and the trial is designed to satisfy a desired level of power as selected by the researcher. A Type II error corresponds to producer's risk, as it corresponds to failing to find a superior experimental treatment when one or more are in

[19] $\beta_{T_k}^C$ will denote the complement of β_{T_k}.

fact superior to the control. Our test will, therefore, be designed to control for a specified one-sided overall Type I error and achieve a desired level of pointwise power.

The minimum number of events among the K experimental treatment groups is a natural test statistic for evaluating the hypotheses in Equation 4.2.1. That is, we will reject the null hypothesis of no difference in efficacy between any of the experimental treatments and the control treatment (in favor of the alternative hypothesis of at least one experimental treatment being superior to the control) if the minimum number of events that occur among the K experimental treatment groups is adequately small.

To determine the Type I error for our test, we need to determine the probability that we will reject the null hypothesis (i.e. find the minimum number of events among the experimental treatment groups to be sufficiently small, say less than or equal to a value m) assuming that the null hypothesis in Equation 4.2.1 is in fact true[20]. Under the null hypothesis, we have a balanced negative multinomial distribution since the event incidence rate is equal to the common value i_C in all of the groups under study[21]. Hence, to compute $P(\min(D_1, \dots, D_K) \leq m)$ under the null hypothesis, we could utilize the formula for the order statistics of a balanced NMD provided by Equation 3.2.2 by setting the index j equal to 1. However, knowing that conditional on $N_C = t$, the number of person years it takes the control group to reach d_C events, the D_k are independent of one another and that the distribution of N_C is $Gamma\left(d_C, \frac{1}{i_C}\right)$, it is simple to calculate the result directly as follows:

$$P(\min(D_1, \dots, D_K) \leq m|t) = 1 - P(\min(D_1, \dots, D_K) > m|t)$$

[20] Appendix C provides a proof that our testing procedure is conservative with respect to Type I error when one or more of the experimental treatments are inferior to the control under the null hypothesis.

[21] Substituting i_C for $i_1, i_2, \dots, i_k$ in Equation 2.4.1 readily shows that the distribution is balanced under the null.

$$= 1 - [P(D_1 > m|t) \cdots P(D_K > m|t)]$$

$$= 1 - \{[1 - P(D_1 \leq m|t)] \cdots [1 - P(D_K \leq m|t)]\}$$

$$= 1 - [1 - P(D_1 \leq m|t)]^K = 1 - \left[1 - \sum_{s=0}^{m} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^K$$

$$\Rightarrow P(\min(D_1, \dots, D_K) \leq m) = \int_0^\infty f(t) \left\{1 - \left[1 - \sum_{s=0}^{m} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^K\right\} dt$$

$$= \int_0^\infty f(t)dt - \int_0^\infty f(t) \left[1 - \sum_{s=0}^{m} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^K dt$$

$$= 1 - \int_0^\infty \frac{(i_C t)^{d_C} t^{-1} e^{-i_C t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-i_C t} \frac{(i_C t)^s}{s!}\right]^K dt$$

Via the same change of variables as in Equation 3.2.1, it can be shown that the integral above is invariant to the value of i_C; in future computations, we will omit mention of this change of variables. Consequently, we have

$$P(\min(D_1, \dots, D_K) \leq m) = 1 - \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-t} \frac{t^s}{s!}\right]^K dt \qquad (4.2.2)$$

Thus, to conduct a trial with a specified one-sided overall Type I error of α_{ovr} given the value of d_C, we must find the critical value m such that Equation 4.2.2 is as close to α_{ovr} as possible without exceeding this value[22]. Due to the discrete nature of the test statistic, it is usually not possible to exactly obtain the specified Type I error. Rather, the nominal Type I error will generally exceed the true Type I error achieved.

[22] This assumes that such a value of m exists given the value of d_C. For small values of d_C, taking $m = 0$ may exceed the nominal Type I error, which also implies that values of m greater than 0 will exceed the nominal Type I error as Equation 4.2.2 is clearly an increasing function in m.

To compute the pointwise power for our test, we must determine the probability that the number of events D_k in a given experimental treatment group is small enough assuming that the incidence rate of events in the experimental treatment group is r_k times as great as that in the control group. The computation is as follows:

$$P(D_k \leq m|t) = \sum_{s=0}^{m} e^{-r_k i_C t} \frac{(r_k i_C t)^s}{s!}$$

$$\Rightarrow P(D_k \leq m) = \int_0^\infty f(t) \left[\sum_{s=0}^{m} e^{-r_k i_C t} \frac{(r_k i_C t)^s}{s!} \right] dt$$

$$= \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[\sum_{s=0}^{m} e^{-r_k t} \frac{(r_k t)^s}{s!} \right] dt \tag{4.2.3}$$

$$= \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[e^{-r_k t} \frac{(r_k t)^0}{0!} + e^{-r_k t} \frac{(r_k t)^1}{1!} + e^{-r_k t} \frac{(r_k t)^2}{2!} + \cdots + e^{-r_k t} \frac{(r_k t)^m}{m!} \right] dt$$

$$= \int_0^\infty \frac{r_k^0 t^{d_C + 0 - 1} e^{-t(1 + r_k)}}{0! \, \Gamma(d_C)} dt + \int_0^\infty \frac{r_k^1 t^{d_C + 1 - 1} e^{-t(1 + r_k)}}{1! \, \Gamma(d_C)} dt + \cdots + \int_0^\infty \frac{r_k^m t^{d_C + m - 1} e^{-t(1 + r_k)}}{m! \, \Gamma(d_C)} dt$$

$$= \sum_{z=0}^{m} \frac{r_k^z}{z! \, \Gamma(d_C)} \int_0^\infty t^{d_c + z - 1} e^{-t/\left(\frac{1}{1 + r_k}\right)} dt = \sum_{z=0}^{m} \frac{r_k^z}{z! \, \Gamma(d_C)} \Gamma(d_C + z) \left(\frac{1}{1 + r_k} \right)^{d_c + z}$$

$$= \sum_{z=0}^{m} \binom{d_c + z - 1}{z} \left(\frac{1}{1 + r_k} \right)^{d_C} \left(1 - \frac{1}{1 + r_k} \right)^z \tag{4.2.3*}$$

Equation 4.2.3 coincides with our definition of pointwise power as it clearly calculates $P\left(\beta_{T_k}^C\right) = 1 - P(\beta_{T_k})$ for the k^{th} experimental treatment group[23]. Thus, we can use this

[23] As pointwise power equates to $P(D_k \leq m)$, we could have used the fact that the marginal distribution of the random variable D_k is negative binomial with parameters d_C and $\frac{p_0}{p_0 + p_k} = \frac{1}{1 + r_k}$. Equation 4.2.3* is then immediate; however, the form of equation 4.2.3 is appealing as it is consistent with the form of additional equations to be derived in Sections 4 and 5.

formula to find the value of m needed to achieve a desired pointwise power given the value of d_C[24]. We find the value of m such that the resulting pointwise power is greater than or equal to the desired power. Once again, these values will generally not coincide due to the discreteness of the underlying distribution.

We will now demonstrate how to design a clinical trial under Design C based on the above results. To design a trial in which both a specified one-sided overall Type I error α_{ovr} and pointwise power are satisfied, we must find the smallest value of d_c and corresponding critical value m such that $P(\min(D_1, \dots, D_K) \leq m) \leq \alpha_{ovr}$ under the null hypothesis in Equation 4.2.1 and $P(\min(D_1, \dots, D_K) \leq m) \geq pointwise\ power$ for a given value of the rate ratio r (i.e. we must find the smallest values of d_C and m which simultaneously satisfy Equations 4.2.2 and 4.2.3 for given values of α_{ovr} and pointwise power)[25]. To accomplish this objective, the function *Des_Sup* was written in R. *Des_Sup* takes the arguments K (number of experimental treatment groups, i.e. not including the control group), *alpha* (nominal one-sided overall Type I error at which the test of hypothesis is to be conducted), *r* (estimate of the rate ratio of the experimental treatment group to the control group which we wish to detect), and pwr (minimum desired pointwise power of the study). *Des_Sup* returns the number of events d_C to be observed in the control group, the critical value m for the hypothesis test, the true overall Type I error achieved, and the true pointwise power achieved (full code for *Des_Sup* can be

[24] Since Equation 4.2.3 is an increasing function in m, we can always find an appropriate value of m to satisfy the desired pointwise power for a given value of d_C. However, for a given value of d_C it may not be possible to simultaneously satisfy a specified Type I error and pointwise power. We will illustrate how to calculate d_C and m to simultaneously satisfy a desired Type I error and pointwise power in the main text.

[25] Though other combinations of d_C and m will also satisfy the desired Type I error and pointwise power, choosing the smallest such d_C and associated m results in the smallest expected number of person years of follow-up until trial termination. The expected number of person years until trial termination will be discussed in detail in Sections 4.5 and 4.6.

found in Appendix E). For example, suppose we have four new treatments to be compared to a single control with one-sided overall Type I error equal to 0.05 and with pointwise power 0.9, and we anticipate the event incidence in a given new treatment group to be 20% that of the event incidence in the control group. Then,

Des_Sup(4,.05,.2,.9) returns

The number of control group events dc is 18
The critical value m is 6
The true overall Type I error is 0.03944082
The true pointwise power is 0.9088288

Hence, the superiority trial will be designed to proceed until 18 events are observed in the control group, and the global null hypothesis will be rejected if the smallest number of events among the four experimental treatment groups is less than or equal to the critical value of 6. Due to the discrete nature of the underlying probability distribution, the true overall Type I error is 0.03944082, which is less than the nominal value of 0.05. Also, for the same reason, the true pointwise power achieved is 0.9088288, which is larger than the specified desired power of 0.9.

The p-value (i.e. the smallest significance level for which the test statistic falls in the rejection region[26]) for the test of treatment superiority can be found using the R function *Prob*, which is called by the *Des_Sup* routine. *Prob* takes the arguments d_C (number of control group events to be observed), m (an integral value), and K (number of experimental treatment groups) and returns $P(\min(D_1, \dots, D_K) \leq m)$ assuming that the null hypothesis is true. Returning to the above example in which the trial continues until 18 events are observed in the control group and the critical value is 6, suppose that the actual minimum number of events observed among the four experimental treatment

[26] See page 63 of Lehmann and Romano (2005) for additional details regarding the p-value for a hypothesis test.

groups is 3. Then the p-value for the trial is $P(\min(D_1, \ldots, D_4) \leq 3)$ under the null hypothesis, and *Prob(18,3,4)* yields the value 0.002885246. As the p-value is less than the specified nominal significance level of 0.05, we would reject the null hypothesis, which is, of course, the same decision that would be made using the critical value approach (i.e. rejecting the null hypothesis since the observed minimum of 3 events is less than or equal to the critical value of $m = 6$).

The number of control group events d_C, critical value m, true one-sided overall Type I error achieved, and true pointwise power achieved in a superiority trial conducted under Design C are presented in columns 2 and 3 of Table 2 below for each combination of nominal $\alpha_{ovr} = 0.05, 0.025, 0.01, 0.001$, nominal pointwise power $= 0.9, 0.8$, $K = 1, 2, 3, 4, 5$, and rate ratio $r = 0.1, 0.2, 0.5$.

Table 2: Number of control group events d_C, critical value m, true one-sided overall Type I error, true pointwise power, and expected person years until trial termination in a superiority trial conducted under Design C for each combination of nominal $\alpha_{ovr} = 0.05, 0.025, 0.01, 0.001$, nominal pointwise power $= 0.9, 0.8$, $K = 1, 2, 3, 4, 5$, and rate ratio $r = 0.1, 0.2, 0.5$

$$\alpha_{ovr} = 0.05, Pointwise\ Power = 0.9$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	9 2	0.03271484 0.9288088	5.89632 (3.269624)	17.53129 (5.918257)	17.53129 (5.918257)	18 (6)	9 2	0.03271484 0.9288088
2	10 2	0.03630554 0.9112841	9.817874 (3.936484)	22.34993 (6.49316)	29.35211 (9.022392)	30 (9.4868)	10 2	0.01928711 0.9112841
3	12 3	0.04708474 0.9587652	17.57756 (5.000736)	31.65715 (7.405265)	47.59235 (13.39679)	48 (13.8564)	13 3	0.01063538 0.948863
4	13 3	0.03753006 0.948863	22.0497 (5.478836)	37.54096 (7.961629)	64.15335 (17.33056)	65 (18.0278)	13 3	0.01063538 0.948863
5	13 3	0.04528845 0.948863	26.38609 (5.885261)	41.50735 (8.185389)	76.89814 (20.55166)	78 (21.6333)	14 3	0.006362915 0.9377837

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	13 5	0.04812622 0.9347349	11.83331 (4.610837)	25.50253 (6.988482)	25.50253 (6.988482)	26 (7.2111)	13 5	0.04812622 0.9347349
2	16 6	0.04850356 0.9394989	22.27046 (5.82601)	38.46302 (8.201388)	47.49208 (11.47404)	48 (12)	17 6	0.01734483 0.9250825
3	17 6	0.04615454 0.9250825	30.10172 (6.512108)	47.31761 (8.87564)	67.07225 (15.59782)	68 (16.4924)	18 6	0.01132792 0.9088288
4	18 6	0.03944082 0.9088288	37.71551 (7.109988)	56.12341 (9.426021)	88.3142 (19.65142)	90 (21.2132)	18 6	0.01132792 0.9088288
5	18 6	0.04744566 0.9088288	45.10579 (7.625637)	63.15559 (9.825743)	106.1196 (23.45096)	108 (25.4558)	20 7	0.009578645 0.9322597

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	47 31	0.04439121 0.9053749	63.61572 (10.83905)	92.91109 (12.93874)	92.91109 (12.93874)	94 (13.7113)	47 31	0.04439121 0.9053749
2	56 36	0.04321293 0.9002963	114.055 (13.1486)	147.6554 (15.37539)	166.6652 (20.91647)	168 (22.4499)	56 36	0.02350578 0.9002963
3	61 39	0.04598938 0.9033769	165.0625 (14.98616)	200.6055 (17.3534)	242.0524 (29.04706)	244 (31.2410)	63 40	0.01484111 0.9001535
4	63 40	0.04951939 0.9001535	211.3623 (16.46559)	247.5616 (18.70317)	312.4537 (36.45178)	315 (39.6863)	68 43	0.01114898 0.9035303
5	68 43	0.04544912 0.9035303	271.5625 (18.35556)	310.6529 (20.55863)	405.0154 (45.42605)	408 (49.4773)	70 44	0.009405374 0.9006217

$$\alpha_{ovr} = 0.05, Pointwise\ Power = 0.8$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	7 1	0.03515625 0.8397133	3.893692 (2.62485)	12.96303 (5.248598)	12.96303 (5.248598)	14 (5.2915)	7 1	0.03515625 0.8397133
2	8 1	0.03687787 0.8057855	6.659308 (3.247875)	16.56288 (5.830937)	22.43389 (7.665144)	24 (8.4853)	8 1	0.01953125 0.8057855
3	11 2	0.03091334 0.8921663	13.39393 (4.443275)	27.12194 (7.042164)	42.73341 (12.32624)	44 (13.2665)	11 2	0.01123047 0.8921663
4	11 2	0.03969332 0.8921663	16.77876 (4.820904)	30.14127 (7.215474)	53.33771 (15.1744)	55 (16.5831)	11 2	0.01123047 0.8921663
5	11 2	0.04791161 0.8921663	20.04923 (5.14312)	33.1418 (7.472)	63.98474 (18.19563)	66 (19.8997)	12 2	0.006469727 0.8716265

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	10 3	0.04614258 0.8419226	7.851527 (3.740763)	18.87073 (6.040165)	18.87073 (6.040165)	20 (6.3246)	10 3	0.04614258 0.8419226
2	13 4	0.04556792 0.8603581	16.07705 (4.976926)	29.84454 (7.272335)	37.77602 (9.867504)	39 (10.8167)	13 4	0.02452087 0.8603581
3	14 4	0.04154506 0.8317516	21.7872 (5.569165)	36.54956 (7.880927)	53.84652 (13.20749)	56 (14.9666)	14 4	0.01544189 0.8317516
4	15 4	0.03398368 0.8011018	27.31122 (6.102428)	43.15971 (8.438999)	71.32572 (16.46652)	75 (19.3649)	15 4	0.009605408 0.8011018
5	15 4	0.04104036 0.8011018	32.67925 (6.541158)	48.12342 (8.725658)	85.3224 (19.43408)	90 (23.2379)	15 4	0.009605408 0.8011018

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	36 22	0.04347445 0.8120462	45.68418 (9.153486)	69.80581 (10.92584)	69.80581 (10.92584)	72 (12)	36 22	0.04347445 0.8120462
2	43 26	0.04880887 0.8150543	83.56242 (11.21573)	110.5964 (13.0836)	126.366 (17.31243)	129 (19.6723)	45 27	0.02218546 0.810087
3	47 28	0.04825873 0.8053409	120.2685 (12.80759)	149.3175 (14.68967)	184.0617 (23.54779)	188 (27.4226)	49 29	0.01539325 0.8008007
4	52 31	0.04631875 0.8146608	165.6697 (14.61111)	197.3862 (16.53801)	254.8906 (30.93439)	260 (36.0555)	54 32	0.0114913 0.810509
5	54 32	0.0468858 0.810509	204.5753 (15.95763)	237.2623 (17.72838)	317.4233 (37.2762)	324 (44.0908)	56 33	0.009593304 0.8065126

$$\alpha_{ovr} = 0.025, Pointwise\ Power = 0.9$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	10 2	0.01928711 0.9112841	5.937091 (3.339847)	19.35887 (6.264188)	19.35887 (6.264188)	20 (6.3246)	10 2	0.01928711 0.9112841
2	13 3	0.02031856 0.948863	13.0325 (4.585059)	29.5666 (7.389085)	38.57321 (10.45835)	39 (10.8167)	13 3	0.01063538 0.948863
3	14 3	0.0178545 0.9377837	17.673 (5.114218)	35.44439 (7.916469)	55.21599 (14.29124)	56 (14.9666)	14 3	0.006362915 0.9377837
4	14 3	0.02310948 0.9377837	22.09674 (5.537852)	39.45998 (8.196804)	68.96947 (17.84165)	70 (18.7083)	15 3	0.003768921 0.9255533
5	15 3	0.01712671 0.9255533	26.48162 (5.994596)	45.28552 (8.672161)	88.24069 (21.74098)	90 (23.2379)	15 3	0.003768921 0.9255533

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	17 6	0.01734483 0.9250825	13.96884 (5.178862)	33.3187 (7.988517)	33.3187 (7.988517)	34 (8.2462)	17 6	0.01734483 0.9250825
2	18 6	0.0215252 0.9088288	22.37659 (5.965809)	42.16297 (8.612329)	53.13649 (11.97718)	54 (12.7279)	18 6	0.01132792 0.9088288
3	21 7	0.01747378 0.9184688	34.38046 (7.083773)	57.20752 (9.755433)	82.86377 (17.2009)	84 (18.3303)	21 7	0.006270476 0.9184688
4	21 7	0.02255745 0.9184688	42.97002 (7.668128)	65.2097 (10.14515)	103.4362 (21.31522)	105 (22.9129)	22 7	0.004065029 0.9031455
5	22 7	0.01819134 0.9031455	51.43399 (8.252204)	75.05559 (10.74458)	129.7626 (25.8041)	132 (28.1425)	22 7	0.004065029 0.9031455

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	56 36	0.02350578 0.9002963	73.80714 (11.89665)	110.7097 (14.10119)	110.7097 (14.10119)	112 (14.9666)	56 36	0.02350578 0.9002963
2	66 42	0.02475517 0.9064897	132.4124 (14.31781)	173.7528 (16.71784)	196.7781 (22.83364)	198 (24.3721)	68 43	0.01114898 0.9035303
3	73 46	0.02276763 0.9068662	193.7204 (16.43456)	238.5947 (18.89399)	289.9412 (31.77569)	292 (34.176)	75 47	0.007085245 0.9041789
4	75 47	0.02479985 0.9041789	247.1244 (17.93168)	292.5698 (20.32476)	372.2272 (39.9049)	375 (43.3013)	77 48	0.005979507 0.9015333
5	80 50	0.02297008 0.9076484	314.3978 (19.91094)	362.5874 (22.17803)	476.6547 (49.45853)	480 (53.6656)	82 51	0.004515245 0.9051912

$$\alpha_{ovr} = 0.025, Pointwise\ Power = 0.8$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	8 1	0.01953125 0.8057855	3.934775 (2.698487)	14.56314 (5.703372)	14.56314 (5.703372)	16 (5.6569)	8 1	0.01953125 0.8057855
2	11 2	0.02147097 0.8921663	9.864372 (3.999337)	24.12886 (6.818107)	32.09112 (9.387985)	33 (9.9499)	11 2	0.01123047 0.8921663
3	12 2	0.01820289 0.8716265	13.43035 (4.494951)	28.86789 (7.372949)	46.3227 (12.73393)	48 (13.8564)	12 2	0.006469727 0.8716265
4	12 2	0.02358442 0.8716265	16.82667 (4.877492)	31.90467 (7.526614)	57.8168 (15.64341)	60 (17.3205)	13 2	0.003692627 0.8498418
5	13 2	0.01687877 0.8498418	20.14621 (5.244107)	36.60701 (8.10451)	74.53605 (19.22603)	78 (21.6333)	13 2	0.003692627 0.8498418

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	13 4	0.02452087 0.8603581	9.930947 (4.32311)	24.85771 (6.937008)	24.85771 (6.937008)	26 (7.2111)	13 4	0.02452087 0.8603581
2	15 4	0.0183677 0.8011018	16.18014 (5.098958)	33.12979 (7.835328)	42.9337 (10.32012)	45 (11.619)	15 4	0.009605408 0.8011018
3	17 5	0.02335044 0.8530007	26.02898 (6.15599)	44.61443 (8.71682)	65.93372 (14.82153)	68 (16.4924)	18 5	0.005311012 0.8275601
4	18 5	0.01933926 0.8275601	32.58649 (6.708971)	52.30302 (9.293028)	86.59155 (18.4784)	90 (21.2132)	18 5	0.005311012 0.8275601
5	18 5	0.02353171 0.8275601	38.9681 (7.198045)	58.2998 (9.613744)	103.821 (21.95478)	108 (25.4558)	19 5	0.003305376 0.8004705

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	45 27	0.02218546 0.810087	55.81427 (10.33226)	87.51736 (12.2425)	87.51736 (12.2425)	90 (13.4164)	45 27	0.02218546 0.810087
2	54 32	0.02164193 0.810509	102.0496 (12.59352)	138.275 (14.60597)	158.992 (19.34488)	162 (22.0454)	54 32	0.0114913 0.810509
3	58 34	0.0218178 0.8026625	144.9614 (14.24042)	183.0177 (16.24331)	227.4298 (26.20047)	232 (30.4631)	58 34	0.008010137 0.8026625
4	63 37	0.02129351 0.8123113	196.3737 (16.08262)	237.1587 (18.05513)	309.3609 (33.92744)	315 (39.6863)	63 37	0.006016488 0.8123113
5	65 38	0.02176037 0.808855	241.3663 (17.44702)	283.1128 (19.37528)	382.825 (40.9684)	390 (48.3735)	67 39	0.004208497 0.8055078

$$\alpha_{ovr} = 0.01, Pointwise\ Power = 0.9$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	14 3	0.006362915 0.9377837	7.983658 (3.957348)	27.4407 (7.423654)	27.4407 (7.423654)	28 (7.4833)	14 3	0.006362915 0.9377837
2	15 3	0.007341652 0.9255533	13.07375 (4.647963)	33.29715 (7.926042)	44.26451 (11.11567)	45 (11.619)	15 3	0.003768921 0.9255533
3	16 3	0.006386439 0.9122111	17.7011 (5.163905)	39.12256 (8.446407)	62.70944 (15.05438)	64 (16)	16 3	0.002212524 0.9122111
4	16 3	0.008366001 0.9122111	22.13765 (5.602838)	43.15541 (8.718219)	78.32144 (18.77424)	80 (20)	16 3	0.002212524 0.9122111
5	19 4	0.006128998 0.9477567	32.81045 (6.681102)	57.4652 (9.731368)	112.6585 (24.94089)	114 (26.1534)	19 4	0.001299739 0.9477567

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	20 7	0.009578645 0.9322597	15.97569 (5.587739)	39.33843 (8.695048)	39.33843 (8.695048)	40 (8.9443)	20 7	0.009578645 0.9322597
2	22 7	0.007881296 0.9031455	25.53829 (6.44582)	51.02433 (9.550633)	64.97177 (13.26258)	66 (14.0712)	22 7	0.004065029 0.9031455
3	24 8	0.009924155 0.9270981	38.54554 (7.505901)	65.26044 (10.46316)	94.96867 (18.59924)	96 (19.5959)	25 8	0.002275692 0.913969
4	25 8	0.008517708 0.913969	48.21195 (8.148895)	76.12009 (11.06351)	123.2021 (23.34782)	125 (25)	25 8	0.002275692 0.913969
5	27 9	0.009062492 0.9347919	63.89735 (9.17552)	93.32036 (11.94657)	160.2432 (29.40114)	162 (31.1769)	27 9	0.001966587 0.9347919

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	70 44	0.009405374 0.9006217	89.90713 (13.317)	138.5446 (15.75112)	138.5446 (15.75112)	140 (16.7332)	70 44	0.009405374 0.9006217
2	82 51	0.008668925 0.9051912	159.9576 (15.89635)	214.5515 (18.55711)	244.4339 (25.45675)	246 (27.1662)	82 51	0.004515245 0.9051912
3	86 53	0.009024783 0.9003806	222.2454 (17.73552)	278.42 (20.36564)	341.5849 (34.35155)	344 (37.0945)	86 53	0.003217405 0.9003806
4	91 56	0.008938622 0.9042405	292.8606 (19.74729)	351.5549 (22.21762)	452.0749 (43.87986)	455 (47.697)	91 56	0.002435871 0.9042405
5*	93 57	0.009273603 0.9020384	357.0994 (21.26332)	416.4126 (23.74615)	554.0885 (53.17317)	558 (57.8619)	96 59	0.001846518 0.9079486

* Design parameters for this row were obtained by substituting 10^3 in the upper limit of the integral for the Type I error formula (see Equation 4.2.2) in the *Des_Sup* code.

$$\alpha_{ovr} = 0.01, Pointwise\ Power = 0.8$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	12 2	0.006469727 0.8716265	5.972554 (3.415234)	22.86864 (6.971402)	22.86864 (6.971402)	24 (6.9282)	12 2	0.006469727 0.8716265
2	13 2	0.007205844 0.8498418	9.904237 (4.071017)	27.57616 (7.522233)	37.48196 (10.06552)	39 (10.8167)	13 2	0.003692627 0.8498418
3	14 2	0.00605492 0.8269907	13.46324 (4.544315)	32.23975 (8.089376)	53.35845 (13.48169)	56 (14.9666)	14 2	0.002090454 0.8269907
4	14 2	0.007943326 0.8269907	16.86934 (4.947104)	35.26515 (8.233229)	66.52979 (16.4656)	70 (18.7083)	14 2	0.002090454 0.8269907
5	14 2	0.009776692 0.8269907	20.17393 (5.277526)	38.27978 (8.444078)	79.68097 (19.61434)	84 (22.4499)	15 2	0.001174927 0.8032494

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	15 4	0.009605408 0.8011018	9.989406 (4.418558)	28.12291 (7.539417)	28.12291 (7.539417)	30 (7.746)	15 4	0.009605408 0.8011018
2	19 5	0.006437142 0.8004705	19.33 (5.647747)	41.90732 (8.887193)	54.71289 (11.6677)	57 (13.0767)	19 5	0.003305376 0.8004705
3	19 5	0.009420385 0.8004705	26.07857 (6.223391)	47.90327 (9.21925)	72.8121 (15.18182)	76 (17.4356)	19 5	0.003305376 0.8004705
4	22 6	0.007021959 0.8265294	37.84301 (7.25511)	63.0637 (10.24375)	106.0673 (20.46976)	110 (23.4521)	22 6	0.001859583 0.8265294
5	22 6	0.008628847 0.8265294	45.27057 (7.77626)	70.10261 (10.6023)	127.2598 (24.35109)	132 (28.1425)	22 6	0.001859583 0.8265294

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	56 33	0.009593304 0.8065126	67.90333 (11.53641)	109.189 (13.70186)	109.189 (13.70186)	112 (14.9666)	56 33	0.009593304 0.8065126
2	65 38	0.009650911 0.808855	120.417 (13.79025)	166.0279 (16.04372)	191.7011 (21.28123)	195 (24.1868)	67 39	0.004208497 0.8055078
3	69 40	0.009864824 0.8022641	169.4451 (15.47111)	216.7916 (17.70869)	271.0437 (28.62121)	276 (33.2265)	72 42	0.003175812 0.8149469
4	74 43	0.009747973 0.811899	226.9781 (17.3317)	276.9251 (19.49638)	363.7499 (36.76967)	370 (43.0116)	76 44	0.002227512 0.8089351
5	78 45	0.008464916 0.8060517	284.1117 (19.03416)	336.754 (21.12187)	459.8289 (44.77405)	468 (52.9906)	78 45	0.001865723 0.8060517

$$\alpha_{ovr} = 0.001, Pointwise\ Power = 0.9$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	20 4	0.0007719398 0.9387666	10.01259 (4.486877)	39.3163 (8.916792)	39.3163 (8.916792)	40 (8.9443)	20 4	0.0007719398 0.9387666
2	21 4	0.0009022344 0.9289592	16.23864 (5.210792)	46.21353 (9.422566)	62.19268 (13.23537)	63 (13.7477)	21 4	0.0004552603 0.9289592
3	22 4	0.0007894748 0.9183452	21.94212 (5.775473)	53.05151 (9.913869)	86.58069 (17.77686)	88 (18.7617)	22 4	0.0002667606 0.9183452
4	23 4	0.0006121234 0.9069417	27.43945 (6.259315)	59.83917 (10.42393)	112.6579 (22.4024)	115 (23.9792)	23 4	0.0001553744 0.9069417
5	23 4	0.0007614876 0.9069417	32.82429 (6.718149)	64.87457 (10.66539)	135.1774 (26.65575)	138 (28.775)	23 4	0.0001553744 0.9069417

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	29 9	0.0008290263 0.910977	19.98139 (6.337554)	56.96278 (10.50991)	56.96278 (10.50991)	58 (10.7703)	29 9	0.0008290263 0.910977
2	32 10	0.0009300359 0.9209501	34.81423 (7.557432)	74.10418 (11.53035)	95.08178 (16.17494)	96 (16.9706)	32 10	0.000470337 0.9209501
3	33 10	0.0008929883 0.9090848	46.85543 (8.344098)	86.90367 (12.12985)	130.3752 (21.47863)	132 (22.9783)	33 10	0.0003030533 0.9090848
4	36 11	0.0006778802 0.9192982	63.66214 (9.389679)	107.0575 (13.2046)	178.178 (28.22763)	180 (30)	36 11	0.0001730006 0.9192982
5	36 11	0.00084194 0.9192982	76.25116 (10.04934)	119.0127 (13.66441)	213.5103 (33.70583)	216 (36)	36 11	0.0001730006 0.9192982

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected Person Years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	107 65	0.0008453577 0.9057528	131.9781 (16.24739)	212.3058 (19.56419)	212.3058 (19.56419)	214 (20.6882)	107 65	0.0008453577 0.9057528
2	116 70	0.0009038727 0.9059773	217.723 (18.63657)	301.3838 (22.08169)	346.2156 (30.31941)	348 (32.311)	116 70	0.00045905 0.9059773
3	120 72	0.0009570337 0.9025216	299.3126 (20.65349)	384.1562 (23.95707)	477.2244 (40.62486)	480 (43.8178)	120 72	0.0003275152 0.9025216
4†	125 75	0.000965024 0.9064813	389.0955 (22.80004)	476.345 (26.02728)	621.8071 (51.86213)	625 (55.9017)	125 75	0.0002497132 0.9064813
5†	129 77	0.0008567777 0.9032685	478.6526 (24.72241)	568.2482 (27.88293)	769.119 (62.66241)	774 (68.1469)	129 77	0.0001781977 0.9032685

† *Des_Sup* did not converge for this row when the infinite upper limit was used in the integral for the Type I error formula; results in this row were obtained by substituting 10^3 in the upper limit in the *Des_Sup* code.

$$\alpha_{ovr} = 0.001, Pointwise\ Power = 0.8$$

$r = 0.1$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	18 3	0.0007448196 0.8824026	7.99608 (3.996223)	34.66714 (8.632338)	34.66714 (8.632338)	36 (8.4853)	18 3	0.0007448196 0.8824026
2	19 3	0.0008483746 0.8660641	13.08907 (4.693224)	40.43011 (9.127256)	55.31585 (12.27038)	57 (13.0767)	19 3	0.0004277229 0.8660641
3	20 3	0.0007235911 0.8488657	17.72216 (5.206056)	46.12221 (9.667516)	77.18071 (16.32121)	80 (17.8885)	20 3	0.0002441406 0.8488657
4	20 3	0.0009592052 0.8488657	22.16348 (5.644948)	50.15802 (9.920079)	96.30801 (20.22392)	100 (22.3607)	20 3	0.0002441406 0.8488657
5	21 3	0.0006808443 0.8308855	26.52995 (6.0715)	55.79915 (10.38724)	120.4843 (24.09199)	126 (27.4955)	21 3	0.0001385808 0.8308855

$r = 0.2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	26 7	0.0006593636 0.827317	16.00201 (5.670755)	49.8514 (10.05425)	49.8514 (10.05425)	52 (10.198)	26 7	0.0006593636 0.827317
2	27 7	0.0008133551 0.8050986	25.56498 (6.498393)	59.49905 (10.69933)	78.26636 (14.05928)	81 (15.5885)	27 7	0.0004106977 0.8050986
3	30 8	0.0006973784 0.8292057	38.5919 (7.573501)	75.76505 (11.64249)	116.6459 (19.55274)	120 (21.9089)	30 8	0.0002359934 0.8292057
4	30 8	0.0009232272 0.8292057	48.23745 (8.194891)	84.79369 (11.99818)	145.5239 (24.02359)	150 (27.3861)	30 8	0.0002359934 0.8292057
5	31 8	0.0007183651 0.8087628	57.73726 (8.786135)	95.4132 (12.5677)	179.504 (28.53651)	186 (33.4066)	31 8	0.0001470384 0.8087628

$r = 0.5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	87 50	0.0009943432 0.8101087	101.9318 (14.27662)	170.5961 (17.16654)	170.5961 (17.16654)	174 (18.6548)	87 50	0.0009943432 0.8101087
2	95 54	0.0009681467 0.800315	169.1548 (16.449)	241.2223 (19.34881)	280.7337 (25.64296)	285 (29.2404)	95 54	0.0004916666 0.800315
3	102 58	0.0009205813 0.8075131	242.5233 (18.6246)	318.3173 (21.52694)	402.3742 (34.92705)	408 (40.398)	102 58	0.0003146735 0.8075131
4‡	106 60	0.0008588862 0.8032139	313.1694 (20.51871)	391.154 (23.34859)	522.2004 (44.01288)	530 (51.4782)	106 60	0.000221663 0.8032139
5	108 61	0.0008950275 0.8011291	381.4415 (22.09594)	460.0734 (24.77339)	638.0635 (52.60929)	648 (62.3538)	108 61	0.0001860506 0.8011291

‡ The Bonferroni values for this row were obtained by substituting 10^3 for the upper limit in the integral for the Type I error formula in the *Des_Sup* code.

As the main interest of researchers is most likely in the probability that experimental treatments which have a given level of efficacy are found to be superior to the control, we chose to power our test using pointwise power. However, other options for power exist as described below.

Another option for power is denoted "partial power" and is defined as the probability that at least one truly superior experimental treatment is discovered, assuming that the alternative hypothesis in Equation 4.2.1 is true. Returning to our example of three experimental treatments labeled A, B, and C, suppose that all of these experimental treatments are superior to the control. Partial power then corresponds to the probability that at least one of A, B, and C are declared statistically superior to the control. Partial power is expressed via the individual Type II errors as

$$Partial\ Power = 1 - P\left(\beta_{T_1} \cap \beta_{T_2} \cap \cdots \cap \beta_{T_K}\right)$$

If only a subset of the K experimental treatments is truly superior, then the number of elements in the expression above should be appropriately reduced; note that this is mainly of theoretical interest as we will not know how many experimental treatments are truly superior to the control in practice. As the number of truly superior experimental treatments increases, partial power increases.

To calculate partial power, we compute $P(\min(D_1, \ldots, D_K) \leq m)$ under the alternative hypothesis in Equation 4.2.1 as follows:

$$P(\min(D_1, \ldots, D_K) \leq m|t) = 1 - \{[1 - P(D_1 \leq m|t)] \cdots [1 - P(D_K \leq m|t)]\}$$

$$= 1 - \left[1 - \sum_{s=0}^{m} e^{-r_1 i_C t} \frac{(r_1 i_C t)^s}{s!}\right]\left[1 - \sum_{s=0}^{m} e^{-r_2 i_C t} \frac{(r_2 i_C t)^s}{s!}\right] \cdots \left[1 - \sum_{s=0}^{m} e^{-r_K i_C t} \frac{(r_K i_C t)^s}{s!}\right]$$

$$\Rightarrow P(\min(D_1, \dots, D_K) \leq m) =$$
$$\int_0^\infty f(t)\left\{1 - \left[1 - \sum_{s=0}^{m} e^{-r_1 i_C t}\frac{(r_1 i_C t)^s}{s!}\right] \dots \left[1 - \sum_{s=0}^{m} e^{-r_K i_C t}\frac{(r_K i_C t)^s}{s!}\right]\right\} dt$$

$$= \int_0^\infty f(t)dt - \int_0^\infty f(t)\left\{\left[1 - \sum_{s=0}^{m} e^{-r_1 i_C t}\frac{(r_1 i_C t)^s}{s!}\right] \dots \left[1 - \sum_{s=0}^{m} e^{-r_K i_C t}\frac{(r_K i_C t)^s}{s!}\right]\right\} dt$$

$$= 1 - \int_0^\infty \frac{(i_C t)^{d_C} t^{-1} e^{-i_C t}}{\Gamma(d_C)}\left[1 - \sum_{s=0}^{m} e^{-r_1 i_C t}\frac{(r_1 i_C t)^s}{s!}\right] \dots \left[1 - \sum_{s=0}^{m} e^{-r_K i_C t}\frac{(r_K i_C t)^s}{s!}\right] dt$$

$$= 1 - \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)}\left[1 - \sum_{s=0}^{m} e^{-r_1 t}\frac{(r_1 t)^s}{s!}\right] \dots \left[1 - \sum_{s=0}^{m} e^{-r_K t}\frac{(r_K t)^s}{s!}\right] dt \qquad (4.2.4)$$

Equation 4.2.4 coincides with the definition of partial power, as it clearly computes $1 - P(\beta_{T_1} \cap \beta_{T_2} \cap \dots \cap \beta_{T_K})$.

Lastly, we define "full power" as the probability that all experimental treatments which are truly superior to the control are found to be superior, assuming that the alternative hypothesis in Equation 4.2.1 is true. For example, suppose treatments A and B are superior to the control, but C is not. Full power would then correspond to the probability that both A and B are found to be statistically superior to the control. Full power is related to the individual Type II errors via

$$Full\ Power = 1 - P\left(\beta_{T_1} \cup \beta_{T_2} \cup \dots \cup \beta_{T_K}\right)$$

As was the case for partial power, if only a subset of the K experimental treatments is truly superior, then the number of elements in the expression for full power should be appropriately reduced. As the number of truly superior experimental treatments increases, full power decreases.

To calculate full power, we compute the probability that the number of events in each new treatment group is less than or equal to the critical value (i.e. is adequately

small in all K new treatment groups). This implies that the maximum number of events among the new treatment groups must be sufficiently small, and so we compute $P(\max(D_1, \dots, D_K) \leq m)$ under the alternative hypothesis in Equation 4.2.1 as follows:

$$P(\max(D_1, \dots, D_K) \leq m|t) = P(D_1 \leq m|t) \cdots P(D_K \leq m|t)$$

$$= \left[\sum_{s=0}^{m} e^{-r_1 i_C t} \frac{(r_1 i_C t)^s}{s!}\right]\left[\sum_{s=0}^{m} e^{-r_2 i_C t} \frac{(r_2 i_C t)^s}{s!}\right] \dots \left[\sum_{s=0}^{m} e^{-r_K i_C t} \frac{(r_K i_C t)^s}{s!}\right]$$

$$\Rightarrow P(\max(D_1, \dots, D_K) \leq m) =$$

$$\int_0^{\infty} f(t)\left[\sum_{s=0}^{m} e^{-r_1 i_C t} \frac{(r_1 i_C t)^s}{s!}\right] \dots \left[\sum_{s=0}^{m} e^{-r_K i_C t} \frac{(r_K i_C t)^s}{s!}\right] dt$$

$$= \int_0^{\infty} \frac{(i_C t)^{d_C} t^{-1} e^{-i_C t}}{\Gamma(d_C)}\left[\sum_{s=0}^{m} e^{-r_1 i_C t} \frac{(r_1 i_C t)^s}{s!}\right] \dots \left[\sum_{s=0}^{m} e^{-r_K i_C t} \frac{(r_K i_C t)^s}{s!}\right] dt$$

$$= \int_0^{\infty} \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)}\left[\sum_{s=0}^{m} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right]\left[\sum_{s=0}^{m} e^{-r_2 t} \frac{(r_2 t)^s}{s!}\right] \dots \left[\sum_{s=0}^{m} e^{-r_K t} \frac{(r_K t)^s}{s!}\right] dt \qquad (4.2.5)$$

This coincides with the definition of full power as it computes $P\left(\beta_{T_1}^C \cap \beta_{T_2}^C \cap \cdots \cap \beta_{T_K}^C\right) = P\left(\beta_{T_1} \cup \beta_{T_2} \cup \cdots \cup \beta_{T_K}\right)^C = 1 - P\left(\beta_{T_1} \cup \beta_{T_2} \cup \cdots \cup \beta_{T_K}\right)$.

Table 3 below compares the pointwise, partial, and full power achieved for given values of d_C and m. To generate this table, *Des_Sup* was used to determine d_C and m for a trial designed to satisfy a one-sided overall Type I error of 0.05 and a pointwise power of 0.9. The corresponding partial power and full power were then obtained by substituting the values of d_C and m into Equations 4.2.4 and 4.2.5, respectively. Values are reported for up to five truly superior experimental treatment groups and for rate ratios of 0.1, 0.2, and 0.5 (the indicated rate ratio is assumed to be the same for all experimental treatment groups under study).

Table 3: Comparison of the values of pointwise, partial, and full power for a test of superiority designed to satisfy a one-sided overall Type I error of 0.05 and a pointwise power of 0.9

$r = 0.1$				
Number of Superior Experimental Treatments	d_C and m	Pointwise Power	Partial Power	Full Power
1	9 2	0.9288088	0.9288088	0.9288088
2	10 2	0.9112841	0.9883564	0.8342119
3	12 3	0.9587652	0.9996792	0.8849313
4	13 3	0.948863	0.9999324	0.8194267
5	13 3	0.948863	0.9999808	0.7827826

$r = 0.2$				
Number of Superior Experimental Treatments	d_C and m	Pointwise Power	Partial Power	Full Power
1	13 5	0.9347349	0.9347349	0.9347349
2	16 6	0.9394989	0.9925503	0.8864476
3	17 6	0.9250825	0.9977752	0.8048559
4	18 6	0.9088288	0.9990017	0.7110355
5	18 6	0.9088288	0.9996498	0.6620089

$r = 0.5$				
Number of Superior Experimental Treatments	d_C and m	Pointwise Power	Partial Power	Full Power
1	47 31	0.9053749	0.9053749	0.9053749
2	56 36	0.9002963	0.9758391	0.8247536
3	61 39	0.9033769	0.9917413	0.770829
4	63 40	0.9001535	0.9999996	0.714471
5	68 43	0.9035303	≈1	0.6810754

Note: The indicated rate ratio applies to all experimental treatment groups under study

4.3: Example of Applying Design C to a Real-World Clinical Trial

In this subsection, we will use data collected from a clinical trial in which multiple experimental influenza vaccines were compared to a single control vaccine to demonstrate the practical implementation of the Design C methodology. Influenza virus infections can lead to respiratory illness, morbidity, and death among both very young and very old persons, as well as among those presenting with comorbidities. Seasonal infection and pandemic influenza is largely controlled via prophylactic vaccination. Such vaccines are usually derived from viruses proliferated in hen eggs; however, the supply of eggs is limited, making production difficult when demand increases unexpectedly. To address this issue, mammalian cell lines have been suggested as alternative culture systems (Frey et al., 2010).

Clinical trial NCT00630331, a randomized, placebo-controlled, observer-blind trial, investigated the efficacy of cell culture-derived influenza vaccine (CCIV) and egg-derived trivalent inactivated vaccine (TIV) compared to a placebo (PBO) in preventing laboratory-confirmed influenza illness in healthy adults during the 2007-2008 influenza season. The study was designed to enroll 11,700 participants, who were equally randomized to the three treatment groups. This sample size was determined based upon individual comparison of each vaccine to the placebo. For a vaccine efficacy of 70%, a one-sided Type I error of 0.0125, and an estimated influenza attack rate of 3%, there was 92% power to reject the null hypothesis that the vaccine efficacy was $\leq$40% for each vaccine. Among the efficacy per protocol population[27] of 11,257 participants, a total of 231 influenza cases occurred; 42 cases among 3,776 subjects in the CCIV group, 49 cases among 3,638 subjects in the TIV group, and 140 cases among 3,843 subjects in the PBO

[27] See Frey et al. (2010) for the definition of the efficacy per protocol population.

group. This corresponds to a CCIV efficacy of 69.5% and a TIV efficacy of 63.0%. The efficacy of each vaccine was highly significant in comparison to the placebo, and both exceeded the Center for Biologics Evaluation and Research vaccine efficacy criteria (Frey et al., 2010).

To conduct this trial using the methodology of Design C, we take the design parameters $K = 2$, one-sided overall Type I error equal to 0.025 (since the individual Type I errors in the trial were constrained at 0.0125, we take our global Type I error to be $2 \times 0.0125 = 0.025$), and pointwise power equal to 0.9. Since the trial was designed assuming a vaccine efficacy of 0.7, we take $r = 0.3$. *Des_Sup(2,.025,.3,.9)* returns

The number of control group events dc is 27
The critical value m is 12
The true overall Type I error is 0.02240684
The true pointwise power is 0.9049494

Hence, under Design C and using pointwise power, the trial would terminate once 27 events are observed in the placebo group, and the null hypothesis of no difference in efficacy between the experimental vaccines and placebo would be rejected if the minimum number of events among the CCIV and TIV groups is less than or equal to 12.

Alternatively, anticipating that both the CCIV and TIV vaccines would be superior to the placebo, investigators may power the study to find both experimental vaccines superior to the placebo, and hence would use the full power formula

$$P(\max(D_1, D_2) \leq m) = \int_0^{\infty} \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[\sum_{s=0}^{m} e^{-r_1 t} \frac{(r_1 t)^s}{s!} \right] \left[\sum_{s=0}^{m} e^{-r_2 t} \frac{(r_2 t)^s}{s!} \right] dt$$

for power computations. The R code for *Des_Sup* can be easily modified to accommodate the definition of full power; taking $r_1 = r_2 = 0.3$, we find the following design parameters:

The number of control group events dc is 31
The critical value m is 15
The true overall Type I error is 0.02439077
The true full power is 0.9096288

For completeness, we also present the resulting design parameters when partial power is used, which corresponds to powering the study to detect at least one truly superior experimental vaccine. In this case, we use

$$P(\min(D_1, D_2) \leq m) = 1 - \int_0^{\infty} \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right] \left[1 - \sum_{s=0}^{m} e^{-r_2 t} \frac{(r_2 t)^s}{s!}\right] dt$$

for the power computations. After modification of the *Des_Sup* code to account for partial power and again taking $r_1 = r_2 = 0.3$, we find

The number of control group events dc is 21
The critical value m is 8
The true overall Type I error is 0.02284066
The true partial power is 0.933896

4.4: Comparison of the Exact Design C Method to the Bonferroni Procedure

Thus far we have considered global hypotheses, that is, we have used information from all study arms simultaneously to determine whether at least one experimental treatment is superior to the control treatment. It may be of interest, however, to compare each experimental treatment to the control individually to identify which treatments (if any) are superior to the control. If we were to individually conduct these K tests, the familywise error rate (the probability of making at least one false rejection among the family of K tests) would become inflated and exceed the specified overall Type I error rate α_{ovr} (Lehmann and Romano, 2005, page 349). To maintain the familywise error rate at α_{ovr} despite the multiple comparisons, the Bonferroni procedure can be used. The Bonferroni procedure conducts each individual test at significance level $\alpha_{ind} = \alpha_{ovr}/K$

to conservatively control the familywise error rate (see Lehmann and Romano (2005), pages 348-350 for details).

We can use the function *Des_Sup* to evaluate the experimental treatments under the Bonferroni procedure. To do so, we set the number of treatment groups equal to one[28] and the Type I error equal to α_{ovr}/K. For example, suppose there are three new treatments being compared to a control treatment and we wish to conduct the trial at significance level $\alpha_{ovr} = 0.05$, pointwise power = 0.8, and for a rate ratio in Equation 4.2.3 of $r = 0.5$. Using the exact Design C method of Section 4.2, we would use *Des_Sup(3,.05,.5,.8)* to find that the trial continues until 47 events are observed in the control group, and the critical value is 28, but, under the Bonferroni procedure, we would use *Des_Sup(1,.05/3,.5,.8)*, which yields 49 events in the control group and a critical value of 29. The difference between the function calls is that for *(3,.05,· ,·)* the calculation is for an overall Type I error of 0.05 when comparing three groups to a control, while for *(1,.05/3, ·, ·)* the calculation is made for an individual Type I error of 0.05/3, which, by the Bonferroni method, is conservative for an overall Type I error of 0.05 for comparing three groups to one control. The cost in using the conservative Bonferroni approach in this example, which requires 49 events in the control group with rejection at ≤29 events in the new treatment group to obtain 0.8 power when $r = 0.5$ for comparing one new treatment to a control at $\alpha_{ind} = 0.05/3$, is an extra two events for the control group and one event for each new treatment arm compared to 47 events in the control group and rejection if ≤28 events occur in at least one new treatment group if the exact calculation is used to compare three new treatment groups to one control at

[28] Since the number of experimental treatment groups is set to one, Equation 2.4.1 reduces to a negative binomial distribution, which is used for testing under the Bonferroni procedure.

$a_{ovr} = 0.05$. Values of d_C, m, true individual Type I error, and true pointwise power obtained under the Bonferroni procedure are included in columns 8 and 9 of Table 2 for comparison with those in columns 2 and 3 which, as previously discussed, were obtained under the exact Design C methodology.

Notice in the example above that both the number of control events to be observed and the critical value is larger under the Bonferroni design; this highlights the fact that there are sometimes considerable savings in terms of the number of events to be observed when the exact method is used compared to the Bonferroni procedure. Thus, though the Bonferroni method may be simpler than the exact method, the disadvantage is that a greater number of control events and/or a greater number of events in the experimental treatment arms imply that it will take a greater number of person years of follow-up for the trial to terminate, which translates to increased study costs. However, for the scenarios presented in Table 2, the losses associated with using the Bonferroni method were not substantial, and, in our discrete setting, values from the Bonferroni and exact method mostly coincided. Regardless of the method used, researchers will be interested in the expected number of person years of follow-up it takes for a trial to terminate. This will be evaluated in the next two subsections.

4.5: Curtailment in Design C

To estimate the expected number of person years of follow-up until trial termination under Design C, we first need to determine the possible mechanisms for discontinuation of subject follow-up and trial termination. Accordingly, in this subsection we define and illustrate uncurtailed and curtailed clinical trials conducted under Design C methodology.

In an uncurtailed trial, recruitment into all treatment arms of the trial and follow-up of all recruited subjects continues until a pre-specified termination point, which, under Design C, occurs when d_C events are observed in the control group. In a curtailed design, recruitment into a given treatment arm (and perhaps follow-up of persons already recruited into that arm) can be discontinued as soon as the ultimate decision is known for the given treatment arm, and the entire trial can be terminated once the ultimate decision is known for all treatment arms. Under Design C, this means that recruitment into each experimental treatment arm can be discontinued once the number of events exceeds the critical value (i.e. once the number of events reaches $m + 1$) as it will no longer be possible to reject the null hypothesis in Equation 4.2.1 based on this experimental treatment group, and the entire trial can be terminated once either (1) all experimental treatment arms reach $m + 1$ events as it will then no longer be possible to reject the null hypothesis, even if we were to continue until all d_C events are observed in the control group, or (2) the control group reaches d_C events at which time the trial is stopped and all remaining active experimental treatment arms are declared superior to the control.

In this dissertation, we will compare the duration of study follow-up under uncurtailed and "fully curtailed" designs. A fully curtailed trial indicates that recruitment is stopped into given treatment arms and these arms have no ongoing follow-up of previously enrolled subjects once they satisfy the conditions for curtailment[29]. We thus assume that either (1) follow-up of subjects in a treatment arm can be discontinued once

[29] Alternatively, it could be argued that follow-up of subjects already recruited into a study must continue for ethical reasons and to increase knowledge of the disease and treatments, perhaps even after d_C events have occurred in the control group. Accordingly, there may be settings in which recruitment of new subjects is stopped, but follow-up of subjects already enrolled in the study (who may not have experienced the outcome of interest) continues, a phenomenon known as "overrunning" (for more information on overrunning, see for example Whitehead, John. "Overrunning and Underrunning in Sequential Clinical Trials." Controlled Clinical Trials, vol. 13, no. 2, Apr. 1992, pp. 106-21).

recruitment into that arm has been terminated, because the outcome of the study for that arm is known, or (2) that the outcome is an immediate short-term binary event of low probability that can be approximated by a Poisson distribution. For the remainder of this dissertation, the term "curtailment" should be interpreted as full curtailment.

Clearly, the expected person years of follow-up in a curtailed design is always less than an uncurtailed design, since there are nonzero probabilities that follow-up of treatment arms (or even the entire study) can be discontinued early, and thus a curtailed design is preferred on this basis. Note that a curtailed design does not change the rejection region for a specified hypothesis test, because curtailment does not alter the ultimate decision made in a trial.

We illustrate the difference in the expected person years of follow-up between an uncurtailed and fully curtailed design via an example. Suppose that two new treatments are being compared to a control in a trial designed to satisfy a one-sided overall Type I error of 0.05, pointwise power = 0.9, and $r = 0.2$. From Table 2, the critical value m for the new treatments is 6, and the number of control group events to observe is $d_C = 16$. In an uncurtailed design, it does not matter how quickly the new treatment groups reach $m + 1 = 7$ events. The entire trial will terminate only once the control group accumulates 16 events, at which time the number of person years of follow-up in each of the new treatment groups is equal to the number of person years it takes the control group to reach 16 events. If we let N_{T_1} and N_{T_2} represent the number of person years of follow-up in the two new treatment arms and N_C the number of person years of follow-up in the

control arm to reach d_C events[30], then the total number of person years of follow-up in this uncurtailed design is $N_{T_1} + N_{T_2} + N_C = N_C + N_C + N_C$ (see (A) in Figure 1 below).

In comparison, consider a curtailed design for the same setting that is stopped early because both new treatment groups reach $m + 1 = 7$ events prior to the control arm reaching $d_C = 16$ events. In this scenario, we know that it will no longer be possible to reject the null hypothesis for either new treatment, even if we were to wait for the control group to reach 16 events, and hence it is futile to continue the study. The total number of person years of follow-up in the control group is then $\max(N_{T_1}, N_{T_2})$, as follow-up in the control arm is curtailed at this time once the study is stopped for futility. The total follow-up across all study arms is then $N_{T_1} + N_{T_2} + \max(N_{T_1}, N_{T_2})$ (see (B1) in Figure 1).

However, if only the first (but not the second) new treatment arm reaches $m + 1 = 7$ events before the control arm reaches 16 events, then the total follow up across all study arms is $N_{T_1} + N_C + N_C$, where N_C is the number of person years needed to reach 16 events in the control arm; N_C is also the follow-up for the second new treatment arm as this arm is discontinued as well as the control arm at N_C person years (see (B2) in Figure 1). Similarly, by symmetry, if only the second (but not the first) new treatment arm reaches 7 events before the control arm reaches 16 events, then the total follow-up across all study arms is $N_C + N_{T_2} + N_C$.

Of course, if the control arm reaches 16 events at N_C person years of follow-up before either new treatment arm reaches 7 events in a curtailed trial, then the total amount

[30] Under a curtailed design, we may not reach d_C events in the control group if the study is terminated early on account of all new treatment groups reaching $m + 1$ events prior to the control group reaching d_C events, and in such cases N_C should be interpreted as the amount of follow-up in the control group at the time of trial termination. It should be well understand by the reader that N_C in an uncurtailed trial will always represent follow-up until d_C events are observed, whereas follow-up may be stopped earlier in a curtailed trial; as such, we do not introduce additional notation to distinguish these interpretations.

of follow-up is $N_C + N_C + N_C$, as all treatment arms are followed for N_C person years (see (B3) in Figure 1).

In the next subsection, we will show how to estimate the expected total number of person years of follow-up for uncurtailed and fully curtailed clinical trials conducted under Design C.

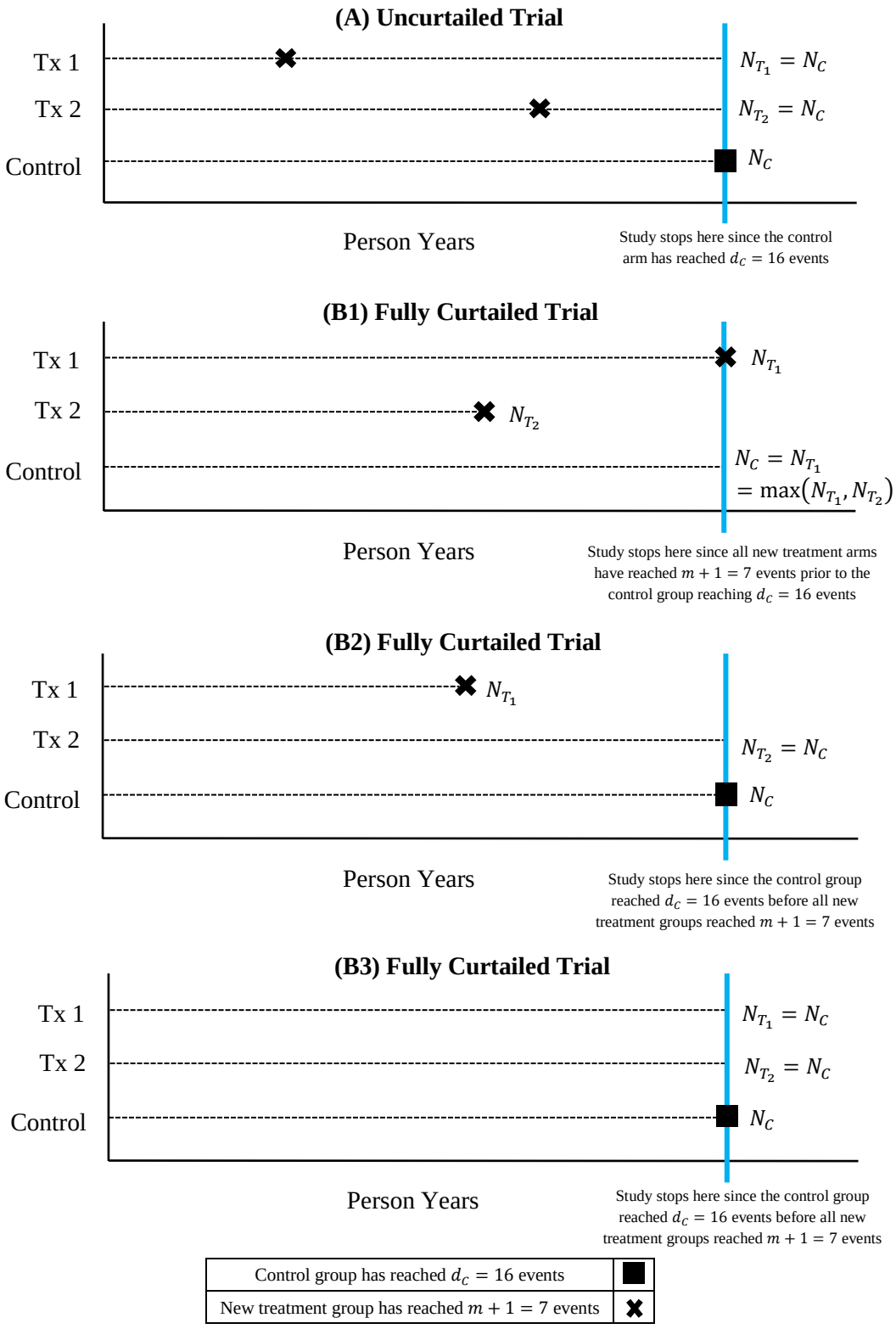


Figure 1: Person years accrued in a superiority trial under an uncurtailed design (A) and for various scenarios under a fully curtailed design (B1-B3) with study parameters $K = 2$, $d_C = 16$, and $m = 6$

4.6: Expected Person Years Under Design C

Researchers and budget personnel will have interest in the expected number of person years of subject follow-up until trial termination, as the longer a study lasts or the larger the number of study participants that must be recruited, the greater the costs to conduct the trial and potentially the shorter the patent life of the agent under study. Thus, in this subsection we discuss formulas and algorithms for estimating the expected number and standard deviation of person years until trial termination under both uncurtailed and fully curtailed designs.

Again, we are working under the assumptions of equal allocation of study subjects to the experimental treatment and control arms (i.e. a $1:1:1 \ldots :1$ allocation ratio) and, as appropriate, immediate discontinuation of follow-up in treatment arms for which the outcome of the study is known. Hence, at any point in time, all active study arms will accrue the same number of person years of follow-up. For example, suppose a trial is to be terminated once the control group reaches d_C events and it takes N_C person years for the control group to accumulate these events. Then all experimental treatment groups still under follow-up at that time will also have incurred N_C person years of follow-up.

The expected number of person years is simple to compute under an uncurtailed design. Under Design C, the trial terminates once the number of events in the control group reaches d_C. Hence, if there are K experimental treatment groups under study, they will also be observed until the number of events in the control group reaches d_C. Thus, the expected number of person years in an uncurtailed design is given by

$$(K+1) \times (expected\ number\ of\ person\ years\ for\ control\ \ to\ reach\ d_C\ events)$$

$= (K + 1) \times E(N_C | d_C)$

We know the number of person years it takes to accumulate d_C events in the control group follows a $Gamma\left(d_C, \frac{1}{i_C}\right)$ distribution. Hence, the expected number of person years for the control arm to reach d_C events is the expected value of this distribution, $\frac{d_C}{i_C}$. So, the expected number of person years until study termination in an uncurtailed trial conducted under Design C is $(K + 1) \times \frac{d_C}{i_C}$. Furthermore, it follows that the variance of the number of person years until termination is given by $(K + 1)^2 \times \frac{d_C}{{i_C}^2}$. For simplicity and without loss of generality, we will take $i_C = 1$ (and so the time it takes the control group to reach d_C events will be assumed to follow a $Gamma(d_C, 1)$ distribution) throughout the remainder of this dissertation.

We now turn our attention to the case of a fully curtailed trial. Suppose we are designing a superiority trial under Design C in which K experimental treatment groups are compared to a single control group. Suppose further that for a given Type I error, power, and values $r_1, r_2, \dots, r_K$ for the rate ratios of event accrual for the experimental treatments, the maximum number of events to observe in the control group is d_C and the critical value is m. A curtailed design makes use of the following rules:

- Follow-up is stopped for any experimental treatment arm that reaches $m + 1$ events prior to the control arm reaching d_C events and H_0 is accepted for all such experimental treatment arms.
- The entire study is stopped if the control arm reaches d_C events (before all experimental treatment arms reach $m + 1$ events) and H_0 is rejected for all experimental treatment arms that have not yet reached $m + 1$ events.

- If all experimental treatment arms reach $m + 1$ events before the control arm reaches d_C events, then the entire study is stopped as H_0 is accepted for all experimental treatment arms.

Let $N_{T_1}, N_{T_2}, \dots, N_{T_K}$ denote the number of person years for experimental treatment arm $1, 2, \dots, K$, respectively, to reach $m + 1$ events, and let N_C denote the number of person years it takes the control arm to reach d_C events[31]. Follow-up for each experimental treatment arm will be terminated once the treatment arm reaches $m + 1$ events, and the number of person years it takes to reach $m + 1$ events in the k^{th} experimental treatment arm follows a $Gamma(m + 1, 1)$ or a $Gamma\left(m + 1, \frac{1}{r_k}\right)$ distribution (corresponding to the null and alternative hypothesis in Equation 4.2.1, respectively, with the value of i_C assumed to be 1). Similarly, follow-up in the control group is terminated once it reaches d_C events, and the number of person years it takes to do so follows a $Gamma(d_C, 1)$ distribution.

There are two settings for which the trial will terminate. In the first, the control group reaches d_C events (thus terminating the trial) before all experimental treatment arms reach $m + 1$ events. In this case, all N_{T_k} for experimental treatment groups which have not reached $m + 1$ events are stopped at N_C person years. In the second setting, all experimental treatment arms reach $m + 1$ events prior to the control group reaching d_C events (thus terminating the trial), so follow-up in the control arm is curtailed at

[31] As was previously mentioned, in a curtailed trial the control group may not reach d_C events if all experimental treatment arms have surpassed the critical value, causing the trial to terminate due to futility. In this case, N_C is interpreted as the amount of follow-up in the control group at the time the trial stops. Similarly, the experimental treatment groups may not reach $m + 1$ events prior to the control group reaching d_C events, at which time the trial is terminated. In this case, $N_{T_1}, N_{T_2}, \dots, N_{T_K}$ are interpreted as the amount of follow-up in the experimental treatment arms at the time of trial stoppage. This is further explained in the main text and is graphically depicted in Figure 2.

$\max_k N_{T_k}$. In both settings, the sum $N_{T_1} + N_{T_2} + \cdots + N_{T_K} + N_C$ represents the total number of person years until the trial terminates. These two settings[32] are depicted graphically in Figure 2 below for a hypothetical study with three new treatments.

[32] In practice, a third setting could arise when d_C events are obtained in the control group at the same time that $m + 1$ events are observed in the final active experimental treatment group. This is possible when a trial has non-continuous follow-up of subjects for the outcome of interest. Clearly, this setting cannot be expressed via the continuous Gamma distributions which characterize subject follow-up. However, in this situation, N_C coincides with $\max_k N_{T_k}$, and so the total number of person years in the study would be identical to that of setting two.

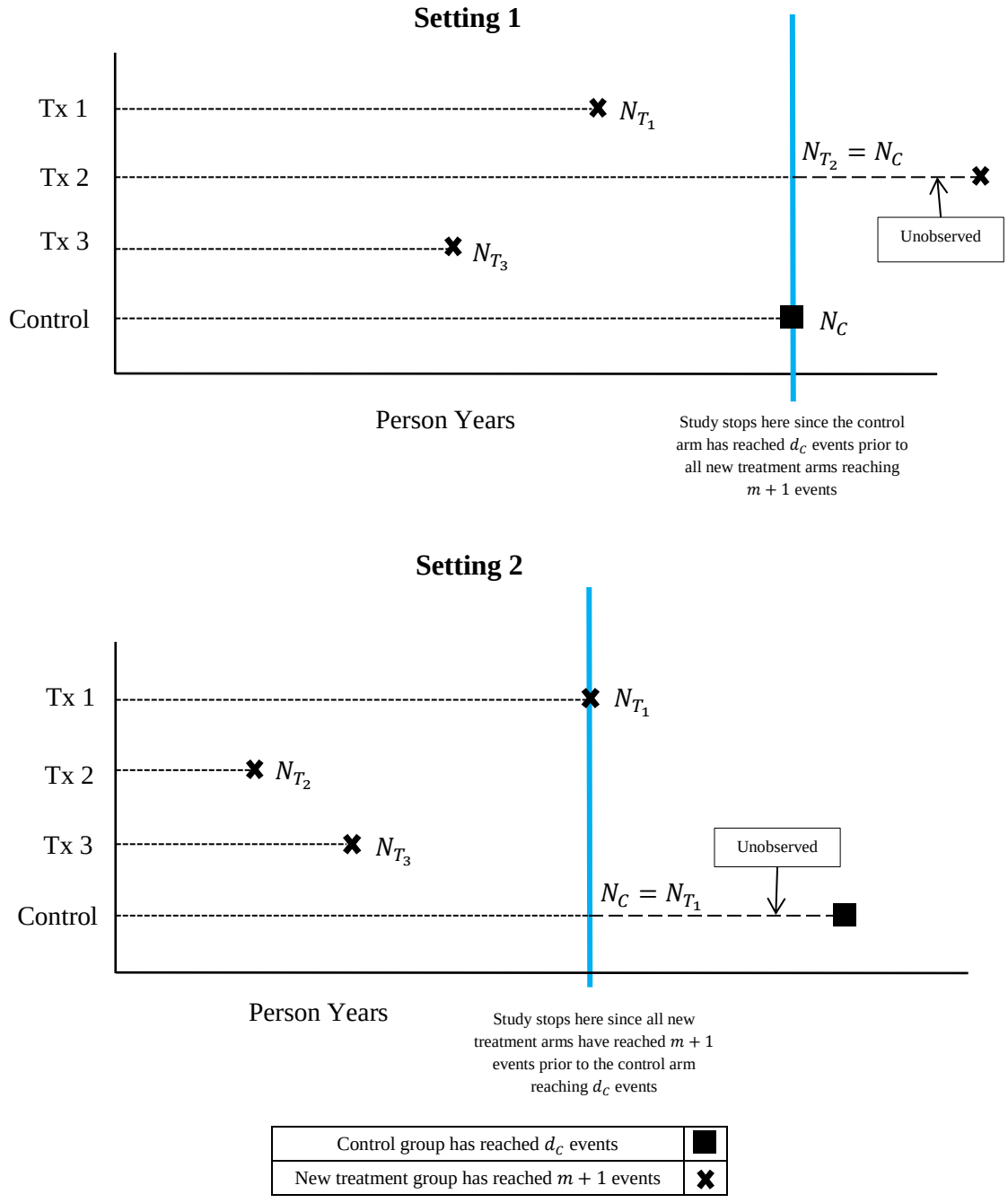


Figure 2: Person years accrued in a fully curtailed superiority trial under two settings: (1) Control arm reaches d_C events prior to all new treatment arms reaching $m + 1$ events; (2) All new treatment arms reach $m + 1$ events prior to the control arm reaching d_C events

To precisely calculate the expected number of person years until trial termination under a fully curtailed design would entail considering all possible ways in which the experimental treatment and control arms could accumulate events and cause the study to terminate, a calculation which is too difficult to explicitly formulate. Therefore, simulation was used to estimate the expected number and standard deviation of person years of follow-up for a fully curtailed design under the null and alternative hypotheses in Equation 4.2.1. The simulation algorithm under the null hypothesis is as follows:

1. Generate K random variables from a $Gamma(m + 1,1)$ distribution to represent the person years for the K experimental treatment groups to exceed the critical value m. Let $\boldsymbol{T_x}$ be a vector containing these K values.
2. Generate a random variable from a $Gamma(d_C, 1)$ distribution to represent the person years for the control group to reach d_C events. Let this random variable be denoted by C.
3. Create the vector $\boldsymbol{T_x^*}$ as follows: for each entry in $\boldsymbol{T_x}$, if the entry is greater than or equal to C, the corresponding entry in $\boldsymbol{T_x^*}$ is set to C. Otherwise, the corresponding entry in $\boldsymbol{T_x^*}$ remains the same as the entry in $\boldsymbol{T_x}$.
4. Create the value C^* as follows: If C is greater than the maximum of the values in $\boldsymbol{T_x}$, set C^* equal to the maximum of the values in $\boldsymbol{T_x}$. Otherwise, set C^* equal to C.
5. Let S equal the sum of all values in $\boldsymbol{T_x^*}$ and C^*.
6. Repeat steps one through five n times, denoting each calculated sum from step five as $S_1, \dots, S_n$, and take the mean and standard deviation of the S_i to estimate the expected number and standard deviation of person years until the trial terminates.

The R function *Null_Time* implements this algorithm and takes the arguments K (number of experimental treatment groups, i.e. not including the control group), dc (number of control group events to be observed), *crit* (the critical value for the test of hypothesis, i.e. m for a superiority trial or w for an inferiority trial[33]), *test* (either "Sup" or "Inf" to specify whether the trial is of treatment superiority or inferiority, respectively), and *sims* (the number of simulations used to estimate the person year values). *Null_Time* also produces a 95% empirical confidence interval for the expected number of person years. Full code for *Null_Time* is provided in Appendix E.

The simulation algorithm to estimate the expected number and standard deviation of person years under the alternative hypothesis in Equation 4.2.1 is as follows:

1. Generate a random variable from a $Gamma\left(m+1, \frac{1}{r_k}\right)$ distribution to represent the person years for the k^{th} experimental treatment group to exceed the critical value m for $k = 1, \ldots, K$. Let $\boldsymbol{T_x}$ be a vector containing these K values.

Steps 2 through 6 follow exactly as in the algorithm presented above for estimation under the null hypothesis.

The R function *Alt_Time* implements this algorithm and takes the arguments dc (number of control group events to be observed), *crit* (the critical value for the test of hypothesis, i.e. m for a superiority trial or w for an inferiority trial), *vec* (a vector of length equal to the number of experimental treatment groups with entries corresponding to the rate ratios of each experimental treatment group, i.e. a vector of the form $c(r_1, r_2, \ldots, r_K)$), *test* (either "Sup" or "Inf" to specify whether the trial is of treatment superiority or inferiority, respectively), and *sims* (the number of simulations used to estimate the person year

[33] Tests of treatment inferiority will be discussed in Section 5.

values). *Alt_Time* also produces a 95% empirical confidence interval for the expected number of person years. Full code for *Alt_Time* is provided in Appendix E.

We next illustrate the use of these algorithms with an example. Suppose a clinical trial is investigating four experimental treatments and is designed to satisfy a one-sided overall Type I error of 0.01, a pointwise power of 0.9, and a hypothesized rate ratio of $r = 0.2$. From Table 2, we know that the critical value for rejecting H_0 for any given experimental treatment arm in this study is $m = 8$, with the stoppage number of events to observe in the control group being $d_C = 25$. *Null_Time(4,25,8,"Sup",100000)* provides an estimated time to termination of 48.21 person years under the null hypothesis in Equation 4.2.1 with an estimated standard deviation of 8.15 person years, based on 100,000 simulations. The associated 95% empirical confidence interval is (35.52, 65.38). We can compare this to an asymptotic 95% confidence interval (based on the normal distribution) for the total number of person years this study will require when the global null hypothesis is true by computing 48.21±1.96×8.15 = (32.24, 64.18) person years. *Alt_Time(25,8,c(.2,.2,.2,.2),"Sup",100000)* yields an estimated time to termination of 123.20 person years and a standard deviation of 23.35 person years under the alternative hypothesis in Equation 4.2.1 when all treatment groups have a rate ratio of 0.2; the 95% empirical confidence interval is (80.98, 171.99). The corresponding asymptotic 95% confidence interval is (77.43, 168.97). If instead the values of the rate ratios are 0.2, 0.5, 0.6, and 0.4 for the four experimental treatment groups, we would use *Alt_Time(25,8,c(.2,.5,.6,.4),"Sup",100000)* to find an estimated number of person years to trial discontinuation of 101.20 with a standard deviation of 14.87, and a 95% empirical

confidence interval of (73.65, 131.82) compared to the asymptotic 95% confidence interval of (72.05, 130.35).

Since the algorithms assume $i_C = 1$, when using the *Null_Time* and *Alt_Time* codes to estimate the expected number and standard deviation of person years until trial termination in practice, the results will need to be multiplied by $\frac{1}{i_C}$. For example, in the hypothetical study of four experimental treatments presented above, the expected number of person years obtained from *Null_Time* was 48.21 with a standard deviation of 8.15. Suppose that the true incidence rate of events in the control arm is 1 event per 10 person years (equivalently, 0.1 events per person year). Then, the estimated expected number of person years until trial termination under the null hypothesis would be 48.21×(1/0.1) = 482.1 person years, and the corresponding standard deviation would be 8.15×(1/0.1) = 81.5 person years.

The hypothetical study above illustrates that for a curtailed superiority trial, the expected number of person years until trial termination under the alternative hypothesis is always greater than under the null hypothesis. Under the null hypothesis, all new treatment groups tend to accumulate events at the same rate as the control group, whereas under the alternative hypothesis, events in the new treatment groups tend to accumulate at rates lower than that of the control group. Hence, it takes longer under the alternative hypothesis for the new treatment arms to reach $m + 1$ events, the time at which follow-up of these treatment arms can be terminated, than under the null hypothesis.

Estimated values for the expected number and standard deviation of person years until trial termination for fully curtailed superiority studies under the null and alternative

hypotheses are provided in columns 4 through 6 in Table 2[34]. In Table 2, *Alt_Time* was computed under two settings: (1) one new treatment group has the rate ratio r specified in the table, and the remaining $K-1$ new treatment groups have rate ratio equal to one (see column 5), and (2) the indicated value of r in the table holds for all experimental treatments under study (i.e. $r_1 = r_2 = \cdots = r_K = r$) (see column 6). Table 2 also contains the expected number and standard deviation of person years until trial discontinuation for an uncurtailed study (with $i_C = 1$) in column 7.

We demonstrate the degree to which full curtailment reduces the expected number of person years in comparison to uncurtailed designs via the graphs in Figure 3 below. The graphs display the expected person years of follow-up in superiority trials for all combinations of the design parameters $\alpha_{ovr} = 0.05, 0.01$, pointwise power $= 0.9, 0.8$, $r = 0.1, 0.2, 0.5$, and $K = 1, 2, 3, 4, 5$. Each graph illustrates the expected number of person years in an uncurtailed study, under the null hypothesis under full curtailment, and under both settings of the alternative hypothesis under full curtailment used to generate columns 5 and 6 in Table 2 (these setting appear in the Figure 3 legend as "Curtailed-Alt (One)" and "Curtailed-Alt (All)", respectively). Plotting the results for both settings of the alternative hypothesis shows that the number of person years under the alternative in a fully curtailed design will depend heavily on the values of the rate ratios of the experimental treatment groups. This is explored further in Figure 4 which illustrates how the expected number of person years under the alternative hypothesis varies with the value of the rate ratio r for the experimental treatment groups. The values of d_C and m used to determine the person years of follow-up in Figure 4 were found using design

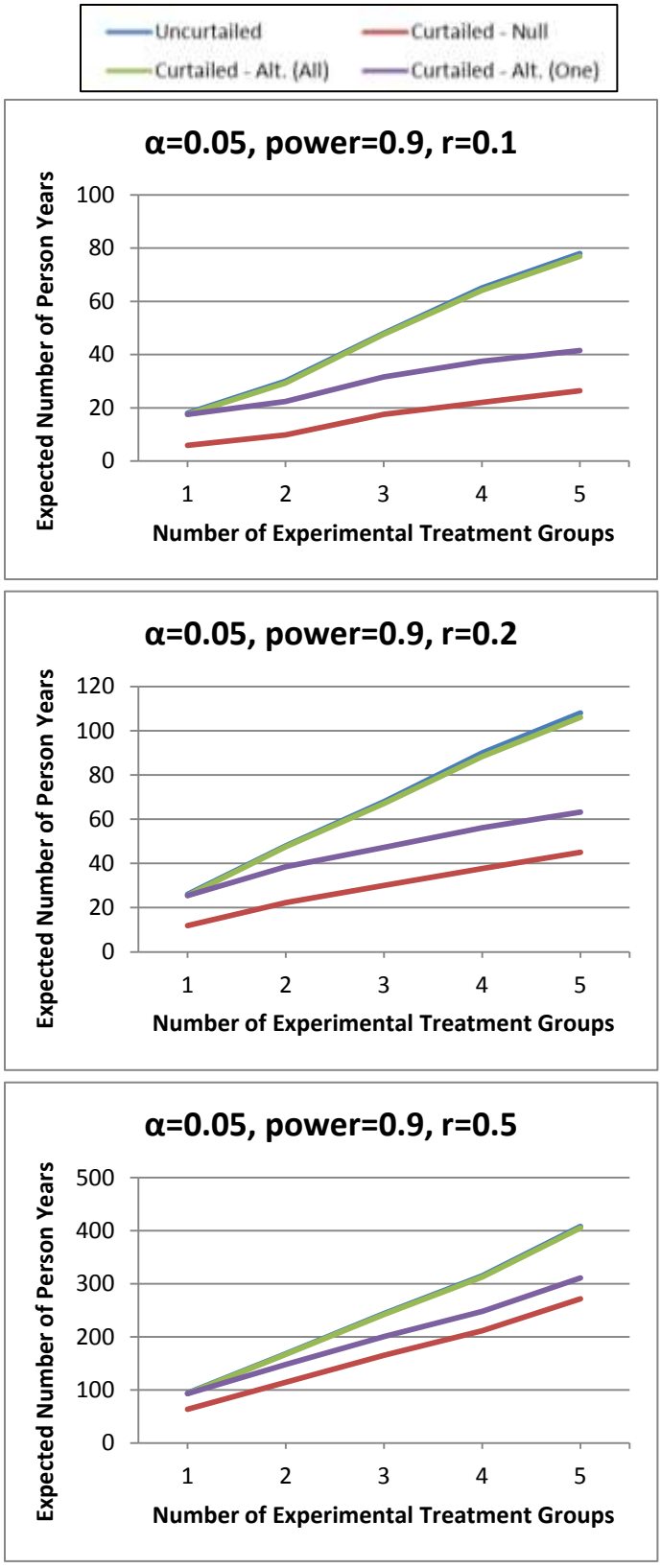
[34] The person year values in columns 4 through 6 of Table 2 were estimated using *Null_Time* and *Alt_Time* with 100,000 simulations; note that these functions return results based on assuming $i_C = 1$.

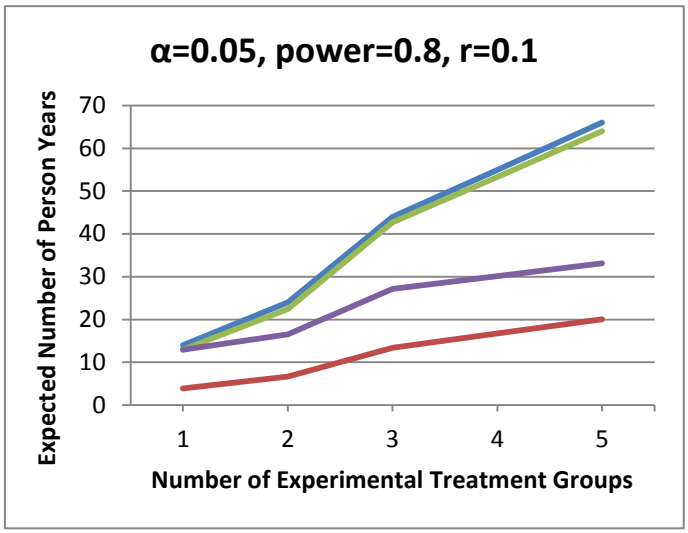
parameters $\alpha_{ovr} = 0.05$, pointwise power $= 0.9$, and $r = 0.1$ in the *Des_Sup* routine, and the indicated value of r in the legend applies to all experimental treatment groups under study.

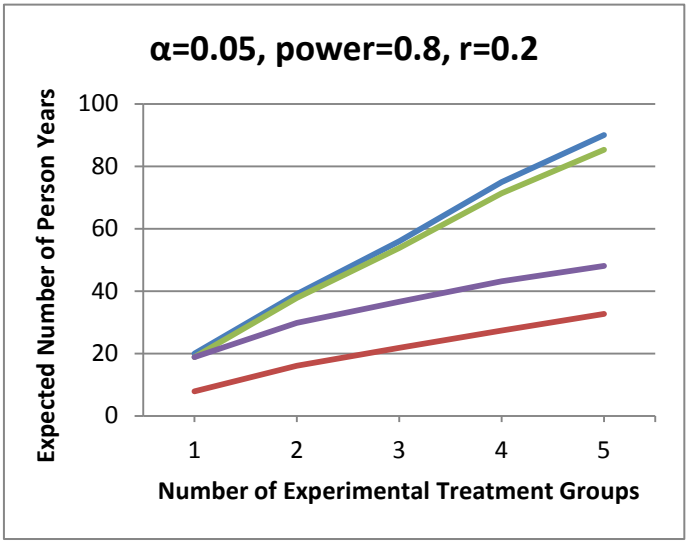
Figure 5 displays the ratio of curtailed to uncurtailed expected person years in a superiority trial for each combination of design parameters in Table 2; in particular, the ratios of the values in column 4 to column 7 (labeled "Curtailed-Null"), column 5 to column 7 (labeled "Curtailed-Alt (One)"), and column 6 to column 7 (labeled "Curtailed-Alt (All)") are plotted in Figure 5.

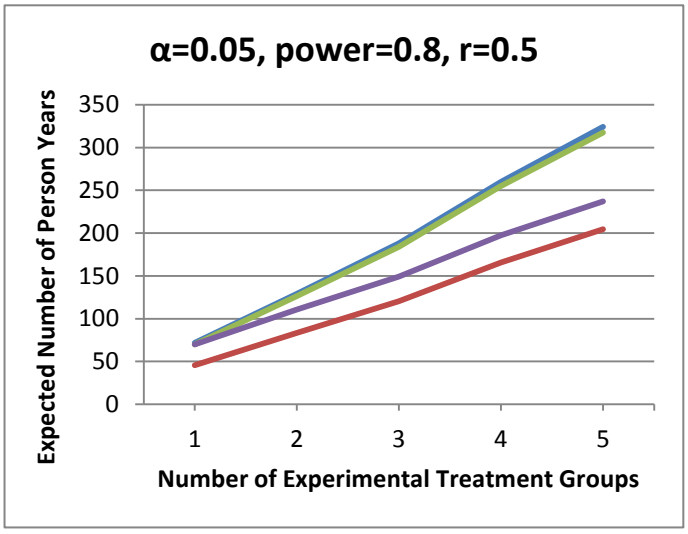
Figures 3, 4, and 5 illustrate that the expected number of person years is reduced the most in a fully curtailed superiority trial when the null hypothesis is true. This implies that pharmaceutical companies can terminate trials most quickly (and thus achieve the greatest possible reduction in study costs) when none of the experimental treatments are superior to the control, and hence, when there is no profit to be made.

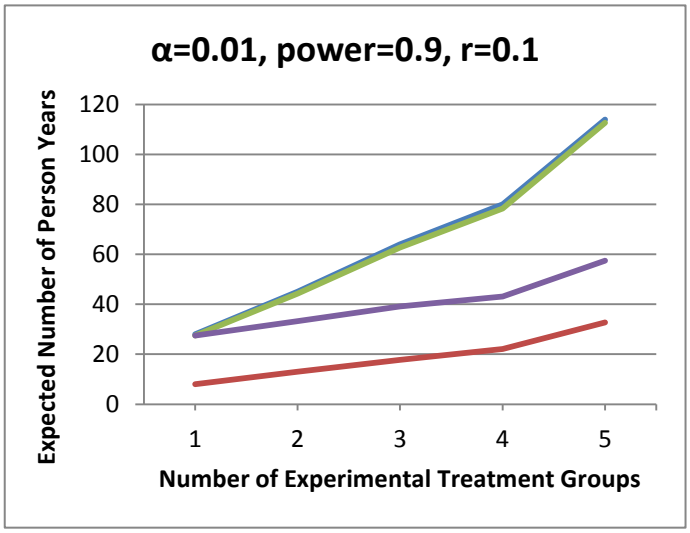
Figure 3: Expected number of person years in a superiority trial for combinations of $\alpha_{ovr} = 0.05, 0.01$, pointwise power $= 0.9, 0.8$, $r = 0.1, 0.2, 0.5$, and $K = 1, 2, 3, 4, 5$

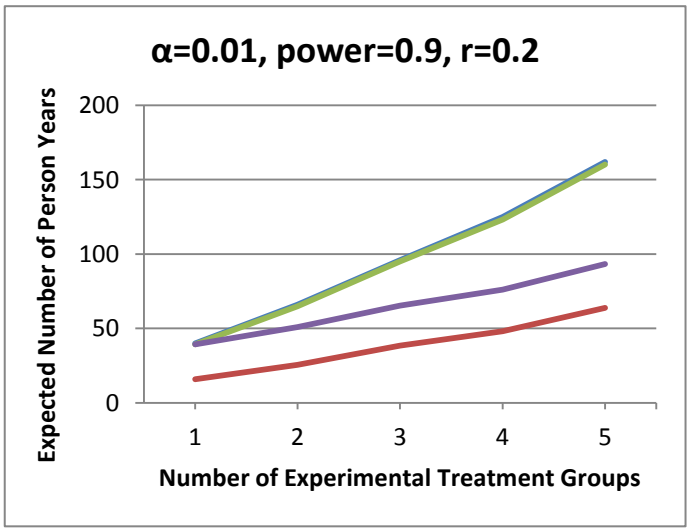


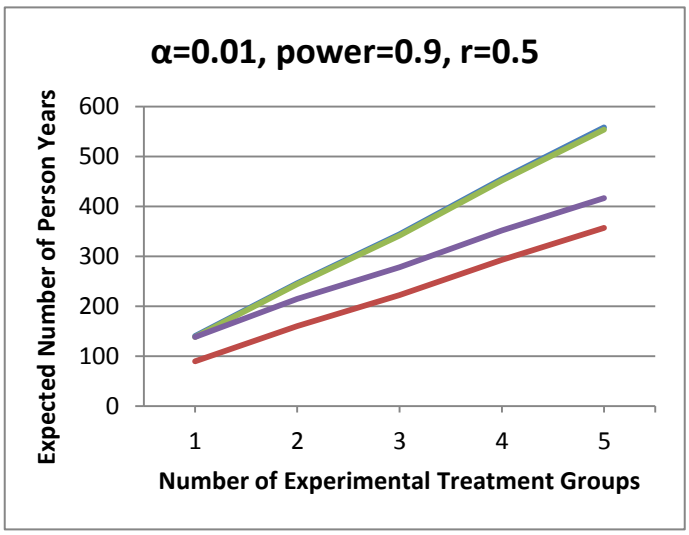

α=0.05, power=0.8, r=0.1
Expected Number of Person Years
70
60
50
40
30
20
10
0
1
2
3
4
5
Number of Experimental Treatment Groups

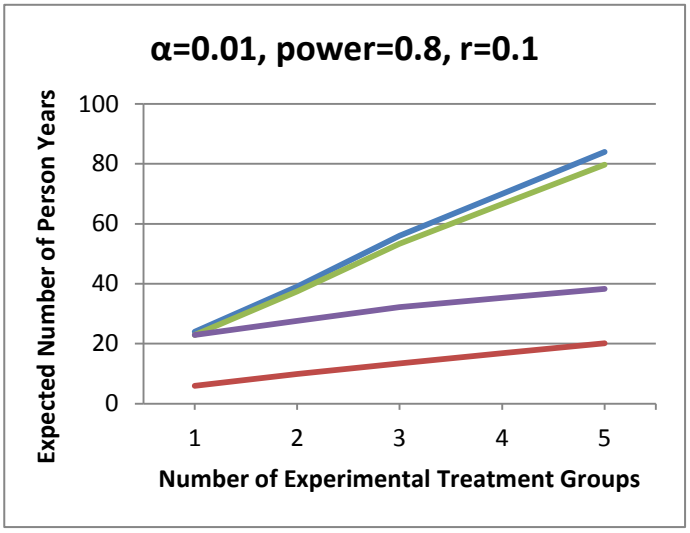

α=0.05, power=0.8, r=0.2
Expected Number of Person Years
100
80
60
40
20
0
1
2
3
4
5
Number of Experimental Treatment Groups

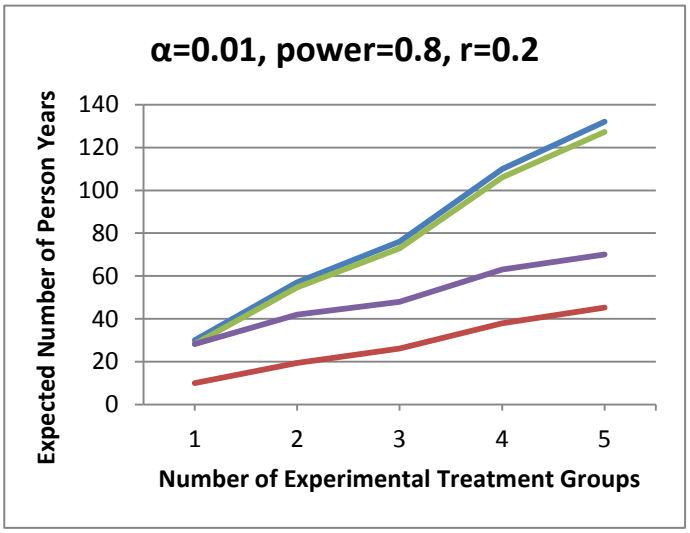

α=0.05, power=0.8, r=0.5
Expected Number of Person Years
350
300
250
200
150
100
50
0
1
2
3
4
5
Number of Experimental Treatment Groups

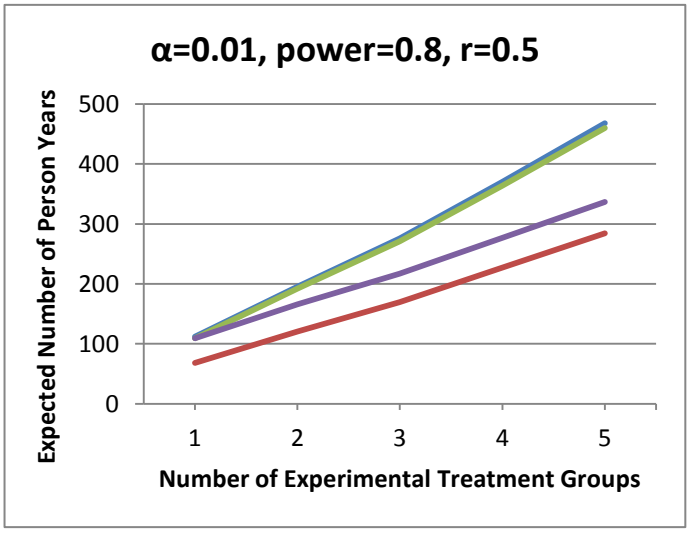

α=0.01, power=0.9, r=0.1
Expected Number of Person Years
120
100
80
60
40
20
0
1
2
3
4
5
Number of Experimental Treatment Groups


α=0.01, power=0.9, r=0.2
Expected Number of Person Years
200
150
100
50
0
1
2
3
4
5
Number of Experimental Treatment Groups


α=0.01, power=0.9, r=0.5
Expected Number of Person Years
600
500
400
300
200
100
0
1
2
3
4
5
Number of Experimental Treatment Groups


α=0.01, power=0.8, r=0.1
Expected Number of Person Years
100
80
60
40
20
0
1
2
3
4
5
Number of Experimental Treatment Groups


α=0.01, power=0.8, r=0.2
Expected Number of Person Years
140
120
100
80
60
40
20
0
1
2
3
4
5
Number of Experimental Treatment Groups


α=0.01, power=0.8, r=0.5
Expected Number of Person Years
500
400
300
200
100
0
1
2
3
4
5
Number of Experimental Treatment Groups

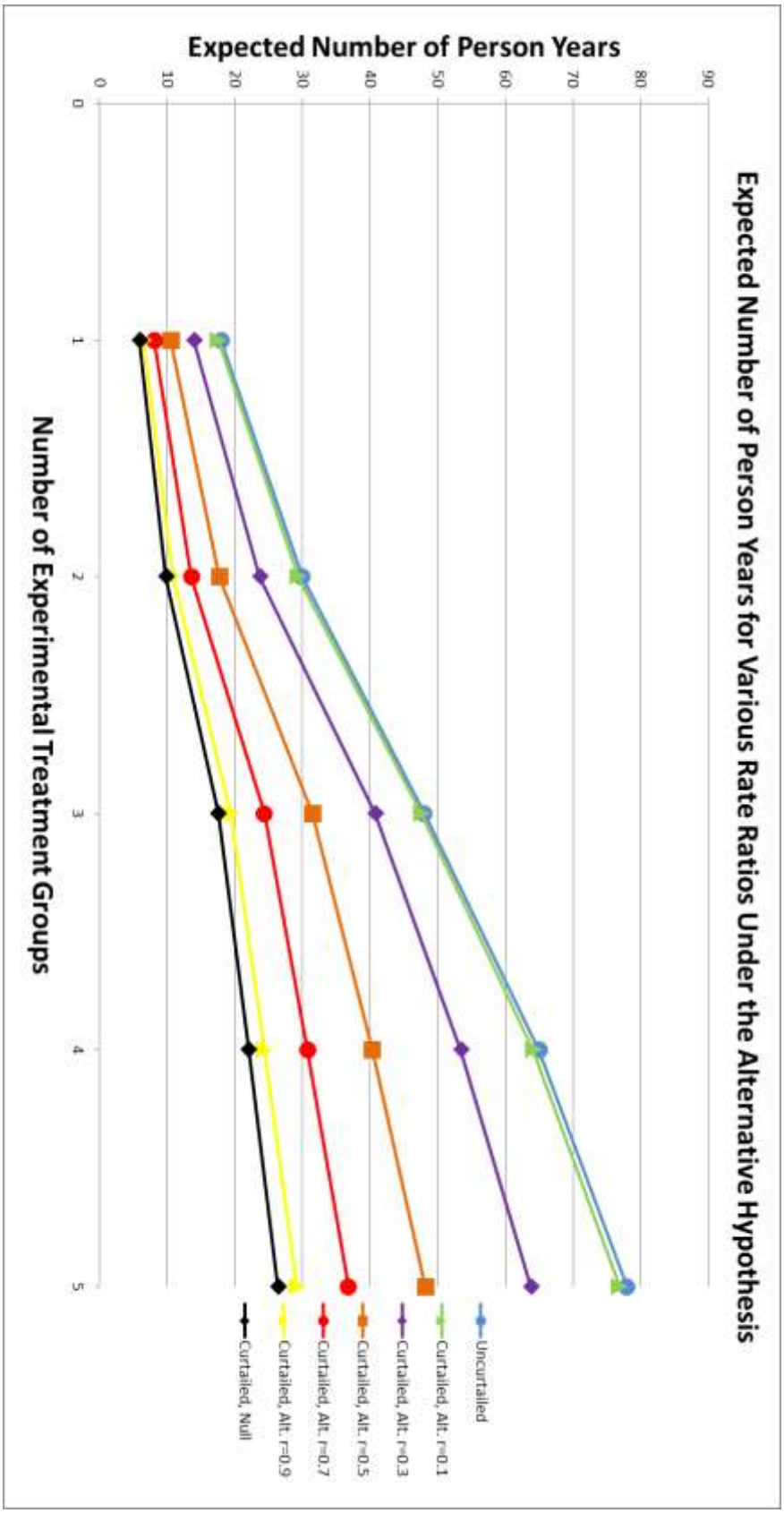
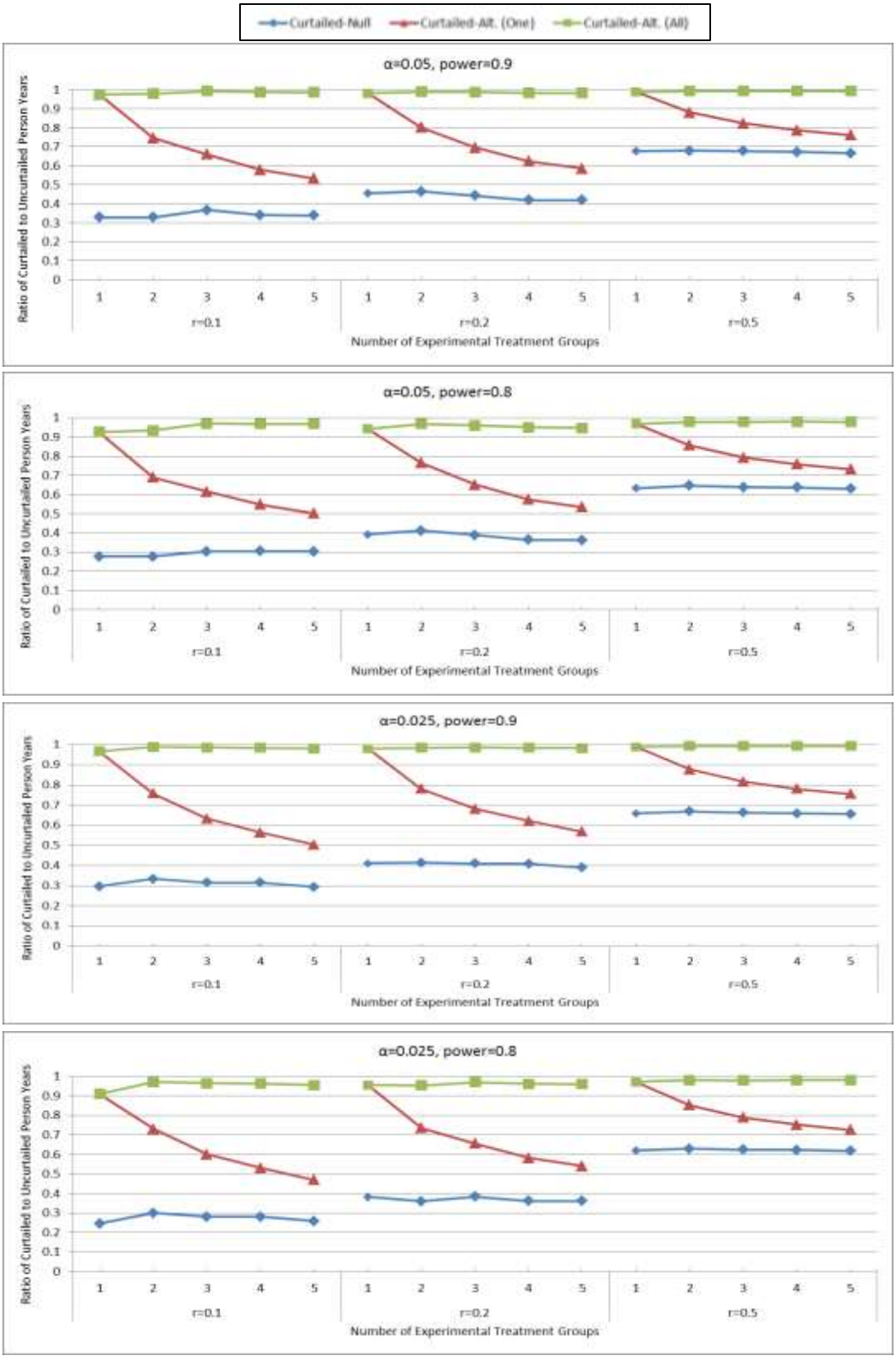
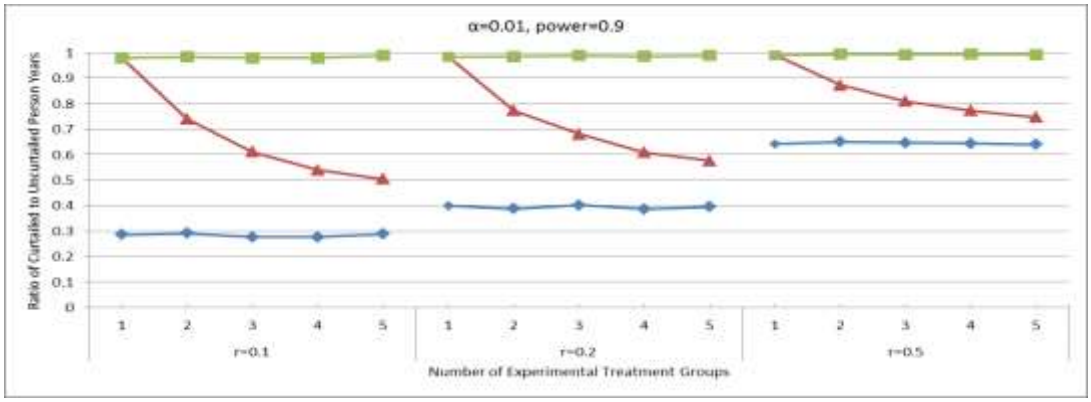
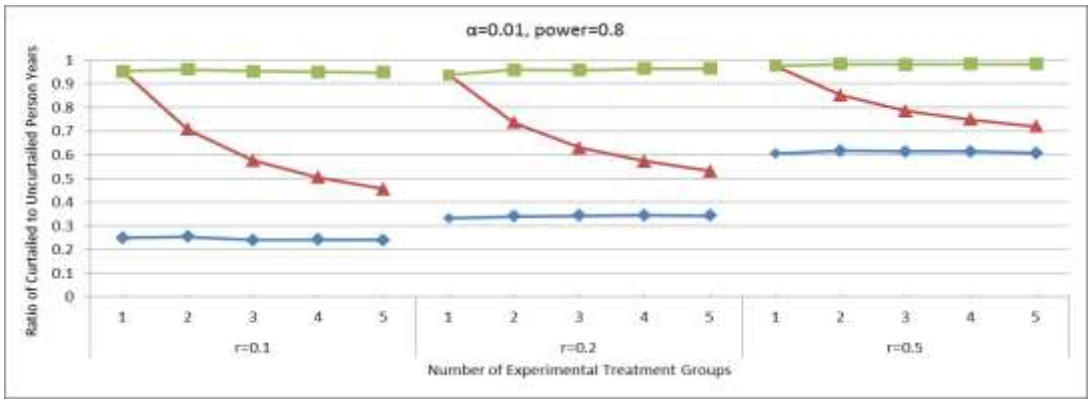


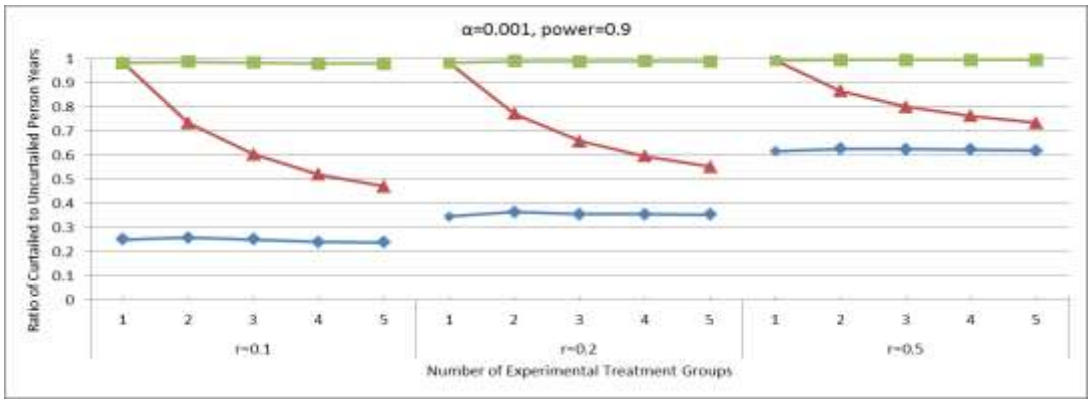
Figure 4: Expected number of person years in a superiority trial for various values of the rate ratios of the experimental treatments with design parameters $\alpha_{ovr} = 0.05$, pointwise power $= 0.9$, $r = 0.1$, and $K = 1, 2, 3, 4, 5$

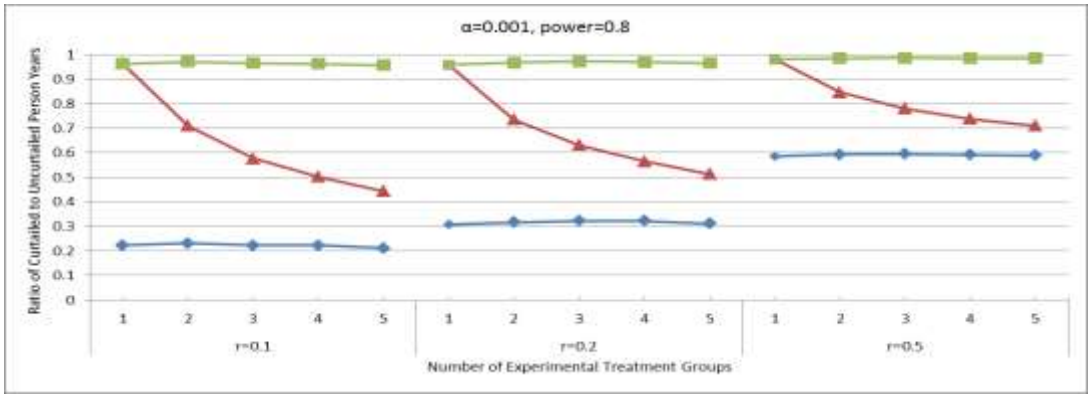
Figure 5: Ratio of fully curtailed to uncurtailed expected person years of follow-up for superiority trials with design parameters given in Table 2




α=0.01, power=0.9
Ratio of Curtailed to Uncurtailed Person Years
r=0.1
r=0.2
r=0.5
Number of Experimental Treatment Groups


α=0.01, power=0.8
Ratio of Curtailed to Uncurtailed Person Years
r=0.1
r=0.2
r=0.5
Number of Experimental Treatment Groups


α=0.001, power=0.9
Ratio of Curtailed to Uncurtailed Person Years
r=0.1
r=0.2
r=0.5
Number of Experimental Treatment Groups


α=0.001, power=0.8
Ratio of Curtailed to Uncurtailed Person Years
r=0.1
r=0.2
r=0.5
Number of Experimental Treatment Groups

SECTION 5: TESTS OF INFERIORITY, TWO-SIDED TESTS OF HYPOTHESES, AND PAIRWISE TESTS OF EXPERIMENTAL TREATMENTS

5.1: Test of Inferiority

In Section 4, we focused on tests of superiority to determine whether at least one experimental treatment is significantly more effective than the control treatment in terms of reducing the incidence of events. In other circumstances, for example early on in the research process, researchers may want to identify experimental treatments that are significantly less effective than the control (i.e. result in significantly more events occurring relative to the control) so that these experimental treatments can be removed from consideration and resources can be reallocated to those more promising agents. To derive a test of treatment inferiority[35], we must define the appropriate counterpart to the alternative hypothesis in Equation 4.2.1; that is, we must define the alternative hypothesis corresponding to at least one experimental treatment being inferior to the control. The hypotheses for an inferiority trial are, therefore, as follows:

$$H_0: i_1 = i_2 = \cdots = i_K = i_C \; versus \; H_a: i_1 = r_1 i_C, i_2 = r_2 i_C, \ldots, i_K = r_K i_C \quad \text{where all } r_k \geq 1 \text{ and at least one of the } r_k \text{ is strictly greater than 1} \qquad (5.1.1)$$

These hypotheses may be relevant during drug discovery when testing of the efficacy of several new agents is conducted. A researcher evaluating several options to improve upon an existing standard of care treatment may want to know if a new treatment being considered is already proven inferior to the standard of care so that further resources are not invested in the new agent or other compounds which have a similar mechanism of action. These hypotheses may also be useful in safety studies where the rare outcome is adverse events which occur during treatment. In this case, acceptance of the null

[35] As was the case for the test of treatment superiority in Section 4, we will assume an equal allocation of study subjects to the experimental and control treatment groups in the derivation of the test of treatment inferiority.

hypothesis indicates an acceptable safety profile of the experimental treatments (i.e. the experimental treatments are not significantly more harmful than the control treatment), and the alternative hypothesis indicates that at least one of the experimental treatments is harmful (i.e. causes too many adverse events in comparison to the control treatment).

It should be noted that for Equation 5.1.1, Type II error corresponds to consumers' risk (as it indicates that one or more experimental treatments are declared to be equally as effective as the control when they are in fact inferior), and Type I error corresponds to producer's risk (as one or more experimental treatments are declared to be inferior when they are in fact equally as effective as the control)[36]. Attempts to manage overall Type I error that make it more difficult to reject the null hypothesis will increase the Type II error (i.e. consumers' risk) and hence may not be desirable. Thus, in practice, investigators will control both Type I and Type II error for an inferiority trial by recruiting an appropriate number of subjects based on sample size calculations and/or by ensuring an adequate amount of subject follow-up.

A natural test statistic for testing the hypotheses in Equation 5.1.1 is the maximum of the $D_k, k = 1,2, \dots, K$. That is, we will reject the null hypothesis of no difference in efficacy between any of the experimental treatments compared to the control treatment (in favor of the alternative hypothesis of at least one experimental treatment being inferior to the control) if the maximum number of events among the K experimental treatment groups is too large, say greater than or equal to a value w (i.e. we reject when too many events occur in at least one experimental treatment group in comparison to the control).

[36] Specifically for a safety study, a Type II error (consumers' risk) occurs when one or more experimental treatments are declared to be safe when they are in fact harmful, and a Type I error (producer's risk) occurs when one or more experimental treatments are declared to be harmful when they are in fact safe.

To calculate the Type I error, we calculate $P(\max(D_1, \dots, D_K) \geq w)$ under the null hypothesis in Equation 5.1.1 as follows:

$P(\max(D_1, \dots, D_K) \geq w|t) = 1 - P(\max(D_1, \dots, D_K) < w|t)$

$= 1 - P(\max(D_1, \dots, D_K) \leq w - 1|t)$

By equation 4.2.5[37], it follows that

$$P(\max(D_1, \dots, D_K) \geq w) = 1 - \int_0^\infty \frac{t^{d_C-1}e^{-t}}{\Gamma(d_C)} \left[\sum_{s=0}^{w-1} e^{-t}\frac{t^s}{s!}\right]^K dt \qquad (5.1.2)$$

Thus, to test the hypotheses in Equation 5.1.1 at a specified one-sided overall Type I error of α_{ovr} and given the value of d_C, we must find the critical value w such that Equation 5.1.2 is as close to α_{ovr} as possible without exceeding this value[38]. Again, due to discreteness, it is usually not possible to exactly obtain the nominal overall Type I error.

Pointwise power (the probability of finding the k^{th} new treatment to be inferior to the control given that it has a rate ratio of r_k) follows readily from Equation 4.2.3 and is given by

$$P(D_k \geq w) = 1 - \int_0^\infty \frac{t^{d_C-1}e^{-t}}{\Gamma(d_C)} \left[\sum_{s=0}^{w-1} e^{-r_k t}\frac{(r_k t)^s}{s!}\right] dt \qquad (5.1.3)$$

or equivalently, by Equation 4.2.3*,

$$P(D_k \geq w) = 1 - \sum_{z=0}^{w-1} \binom{d_c + z - 1}{z} \left(\frac{1}{1+r_k}\right)^{d_C} \left(1 - \frac{1}{1+r_k}\right)^z \qquad (5.1.3^*)$$

---

[37] Throughout Section 5.1, we will make use of several formulas related to the minimum and maximum of $D_1, D_2, \dots, D_K$ which were derived in Section 4.2. In Section 4.2, the value of all rate ratios were assumed to be ≤ 1 as stated in Equation 4.2.1. However, when the results in Section 4.2 are applied here in Section 5.1, the rate ratios in the resulting formulas are assumed to have value ≥ 1 to reflect the assumptions of Equation 5.1.1.

[38] Since Equation 5.1.2 is a decreasing function in w, we can always find a value of w that satisfies the desired Type I error given the value of d_C. However, this value may not satisfy a desired pointwise power (see Equation 5.1.3 for pointwise power in an inferiority study). We will show how to find values of d_C and w that simultaneously achieve a desired Type I error and pointwise power in the main text.

We now show how to design a trial to test the hypotheses in Equation 5.1.1 at a specified one-sided overall Type I error α_{ovr} and which achieves a desired pointwise power. We must find the smallest value d_C and corresponding critical value w such that $P(\max(D_1, \ldots, D_K) \geq w) \leq \alpha_{ovr}$ under the null hypothesis in Equation 5.1.1 and $P(\max(D_1, \ldots, D_K) \geq w) \geq pointwise\ power$ for a given value of the rate ratio r (i.e. we must find the smallest values of d_C and w that simultaneously satisfy Equations 5.1.2 and 5.1.3 for given values of α_{ovr} and pointwise power). To determine d_C and w, the function *Des_Inf* was written in R. This function takes the same arguments as *Des_Sup* (i.e., $K, alpha, r, pwr$) and returns the number of events d_C to be observed in the control group, the critical value w, the true overall Type I error achieved, and the true pointwise power achieved in an inferiority trial conducted under Design C methodology. For example, suppose there are four experimental treatments under study and researchers want to determine if any of them have an unacceptable safety profile, as indicated by causing significantly more adverse events than a control treatment. Researchers aim to detect experimental treatments that cause at least twice as many adverse events as the control group (i.e. $r = 2$). To test the hypotheses in Equation 5.1.1 at a one-sided overall Type I error of 0.05 and to achieve a minimum pointwise power of 0.8, the function *Des_Inf(4,.05,2,.8)* returns the design parameters

The number of control group events dc is 30
The critical value w is 49
The true overall Type I error is 0.04866245
The true pointwise power is 0.8008007

Hence, the study would continue until 30 events are observed in the control group, and the global null hypothesis will be rejected (indicating at least one of the experimental treatments is harmful) if 49 or more adverse events have occurred in any of the

experimental treatment arms. Those treatments that cause 49 or more adverse events would be removed from future consideration as research continues. Full code for the *Des_Inf* function is provided in Appendix E.

Columns 2 and 3 in Table 4 below provide the number of control group events d_C, critical value w, true one-sided overall Type I error, and true pointwise power achieved in an inferiority trial conducted under Design C for each combination of nominal $\alpha_{ovr} = 0.05, 0.025, 0.01, 0.001$, nominal pointwise power $= 0.9, 0.8$, $K = 1, 2, 3, 4, 5$, and rate ratio $r = 10, 5, 2$. The corresponding values obtained under the Bonferroni procedure are provided for comparison in columns 8 and 9.

Table 4: Number of control group events d_C, critical value w, true one-sided overall Type I error, true pointwise power, and expected person years until trial termination in an inferiority trial conducted under Design C for each combination of nominal $\alpha_{ovr} = 0.05, 0.025, 0.01, 0.001$, nominal pointwise power $= 0.9, 0.8$, $K = 1, 2, 3, 4, 5$, and rate ratio $r = 10, 5, 2$

$$\alpha_{ovr} = 0.05, Pointwise\ Power = 0.9$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	3 9	0.03271484 0.9288088	5.898793 (3.276461)	1.751874 (0.5902905)	1.751874 (0.5902905)	6 (3.4641)	3 9	0.03271484 0.9288088
2	3 10	0.03419927 0.9112841	8.923622 (5.023876)	6.903292 (3.391291)	3.06093 (0.7516046)	9 (5.1962)	3 10	0.01928711 0.9112841
3	3 10	0.04653185 0.9112841	11.94335 (6.730296)	9.927413 (5.110737)	4.113392 (0.8706497)	12 (6.9282)	4 13	0.01063538 0.948863
4	4 13	0.03319684 0.948863	19.91971 (9.807395)	17.21691 (7.892682)	6.755959 (1.053144)	20 (10)	4 13	0.01063538 0.948863
5	4 13	0.03904195 0.948863	23.92759 (11.76612)	21.22344 (9.85241)	8.089269 (1.154536)	24 (12)	4 14	0.006362915 0.9377837

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	6 13	0.04812622 0.9347349	11.83748 (4.615357)	5.100476 (1.394228)	5.100476 (1.394228)	12 (4.899)	6 13	0.04812622 0.9347349
2	7 16	0.04644442 0.9394989	20.8838 (7.700184)	17.05177 (5.242304)	9.87437 (1.798618)	21 (7.9373)	7 17	0.01734483 0.9250825
3	7 17	0.04282289 0.9250825	27.92929 (10.3649)	24.29017 (7.921618)	13.99833 (2.109547)	28 (10.583)	7 18	0.01132792 0.9088288
4	7 18	0.03569877 0.9088288	34.90644 (12.9889)	31.44871 (10.55262)	18.39816 (2.495376)	35 (13.2288)	7 18	0.01132792 0.9088288
5	7 18	0.04206158 0.9088288	41.94544 (15.63896)	38.47876 (13.18809)	22.02696 (2.757675)	42 (15.8745)	8 20	0.009578645 0.9322597

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	32 47	0.04439121 0.9053749	63.61826 (10.84462)	46.44982 (6.453644)	46.44982 (6.453644)	64 (11.3137)	32 47	0.04439121 0.9053749
2	36 54	0.04983791 0.9039034	107.6929 (17.54011)	98.4466 (12.57864)	81.98854 (7.775)	108 (18)	37 56	0.02350578 0.9002963
3	40 61	0.04437004 0.9033769	159.789 (24.88359)	150.0521 (19.52604)	123.7223 (9.259815)	160 (25.2982)	41 63	0.01484111 0.9001535
4	41 63	0.04722764 0.9001535	204.6567 (31.45184)	194.9315 (25.97006)	159.4957 (10.45654)	205 (32.0156)	44 68	0.01114898 0.9035303
5	44 68	0.04298241 0.9035303	263.6148 (39.35477)	253.3574 (33.55302)	206.2918 (11.95103)	264 (39.7995)	45 70	0.009405374 0.9006217

$$\alpha_{ovr} = 0.05, Pointwise\ Power = 0.8$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	2 7	0.03515625 0.8397133	3.893 (2.621769)	1.295922 (0.5237544)	1.295922 (0.5237544)	4 (2.8284)	2 7	0.03515625 0.8397133
2	2 8	0.03442004 0.8057855	5.928759 (4.081039)	4.669951 (2.800944)	2.307309 (0.7361318)	6 (4.2426)	2 8	0.01953125 0.8057855
3	2 8	0.04663015 0.8057855	7.908185 (5.411604)	6.66084 (4.154902)	3.103153 (0.8992955)	8 (5.6569)	3 11	0.01123047 0.8921663
4	3 11	0.0346426 0.8921663	14.95992 (8.455921)	13.02681 (6.841925)	5.59241 (1.0836)	15 (8.6603)	3 11	0.01123047 0.8921663
5	3 11	0.04064392 0.8921663	17.94046 (10.16118)	16.01038 (8.53627)	6.694534 (1.214613)	18 (10.3923)	3 12	0.006469727 0.8716265

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	4 10	0.04614258 0.8419226	7.856246 (3.751093)	3.773724 (1.20563)	3.773724 (1.20563)	8 (4)	4 10	0.04614258 0.8419226
2	5 13	0.04336439 0.8603581	14.89835 (6.498715)	12.39962 (4.509587)	7.802649 (1.686832)	15 (6.7082)	5 13	0.02452087 0.8603581
3	5 14	0.03815039 0.8317516	19.94745 (8.730277)	17.61619 (6.778049)	11.13858 (2.12693)	20 (8.9443)	5 14	0.01544189 0.8317516
4	5 14	0.04720239 0.8317516	24.9392 (10.96543)	22.61022 (8.998929)	13.90014 (2.473376)	25 (11.1803)	5 15	0.009605408 0.8011018
5	5 15	0.03580065 0.8011018	29.89399 (13.19443)	27.72576 (11.25984)	17.57759 (3.115025)	30 (13.4164)	5 15	0.009605408 0.8011018

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	23 36	0.04347445 0.8120462	45.6915 (9.167689)	34.90525 (5.466055)	34.90525 (5.466055)	46 (9.5917)	23 36	0.04347445 0.8120462
2	27 43	0.04762805 0.8150543	80.79932 (15.18776)	74.73955 (11.22864)	64.26006 (7.016001)	81 (15.5885)	28 45	0.02218546 0.810087
3	29 47	0.04630678 0.8053409	115.7432 (21.02785)	109.6676 (16.89242)	93.71933 (8.617734)	116 (21.5407)	30 49	0.01539325 0.8008007
4	30 49	0.04866245 0.8008007	149.8836 (27.05042)	143.749 (22.7615)	121.8773 (10.18055)	150 (27.3861)	33 54	0.0114913 0.810509
5	33 54	0.04400853 0.810509	197.7139 (34.09813)	191.0972 (29.49175)	161.2836 (11.9323)	198 (34.4674)	34 56	0.009593304 0.8065126

$$\alpha_{ovr} = 0.025, Pointwise\ Power = 0.9$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	3 10	0.01928711 0.9112841	5.937082 (3.341794)	1.934689 (0.6246614)	1.934689 (0.6246614)	6 (3.4641)	3 10	0.01928711 0.9112841
2	4 13	0.0192862 0.948863	11.96407 (5.908739)	9.24741 (3.971093)	4.027862 (0.8320869)	12 (6)	4 13	0.01063538 0.948863
3	4 14	0.01634412 0.9377837	15.96754 (7.887217)	13.34916 (5.969545)	5.790758 (0.9999513)	16 (8)	4 14	0.006362915 0.9377837
4	4 14	0.02050249 0.9377837	19.9567 (9.902123)	17.33867 (7.976523)	7.231522 (1.123349)	20 (10)	4 15	0.003768921 0.9255533
5	4 14	0.02428145 0.9377837	23.92454 (11.83856)	21.31323 (9.918273)	8.656039 (1.241889)	24 (12)	4 15	0.003768921 0.9255533

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	7 17	0.01734483 0.9250825	13.96953 (5.179042)	6.664143 (1.594557)	6.664143 (1.594557)	14 (5.2915)	7 17	0.01734483 0.9250825
2	7 18	0.02062968 0.9088288	20.94186 (7.819463)	17.46969 (5.377859)	10.98579 (1.935409)	21 (7.9373)	7 18	0.01132792 0.9088288
3	8 20	0.024463 0.9322597	31.99767 (11.1586)	27.93848 (8.511046)	16.46303 (2.273721)	32 (11.3137)	8 21	0.006270475 0.9184688
4	8 21	0.02055981 0.9184688	39.96106 (13.99153)	36.09587 (11.35407)	21.49082 (2.656032)	40 (14.1421)	8 22	0.004065028 0.9031455
5	8 21	0.02443801 0.9184688	48.06447 (16.88735)	44.18032 (14.22741)	25.7329 (2.934329)	48 (16.9706)	8 22	0.004065028 0.9031455

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	37 56	0.02350578 0.9002963	73.80503 (11.89769)	55.34858 (7.038925)	55.34858 (7.038925)	74 (12.1655)	37 56	0.02350578 0.9002963
2	43 66	0.02426523 0.9064897	128.7276 (19.42252)	118.4836 (13.94251)	100.1501 (8.622093)	129 (19.6723)	44 68	0.01114898 0.9035303
3	45 70	0.02454158 0.9006217	179.8053 (26.67894)	169.511 (20.91968)	141.7709 (9.934298)	180 (26.8328)	48 75	0.007085245 0.9041789
4	48 75	0.02372738 0.9041789	239.9857 (34.41193)	229.1444 (28.37823)	189.7654 (11.40086)	240 (34.641)	49 77	0.005979507 0.9015333
5	49 77	0.0241933 0.9015333	293.9446 (41.84614)	283.1032 (35.66927)	233.3704 (12.68077)	294 (42)	52 82	0.004515245 0.9051912

$$\alpha_{ovr} = 0.025, Pointwise\ Power = 0.8$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	2 8	0.01953125 0.8057855	3.933693 (2.694395)	1.455974 (0.5693162)	1.455974 (0.5693162)	4 (2.8284)	2 8	0.01953125 0.8057855
2	3 11	0.02026146 0.8921663	8.946299 (5.088623)	7.013559 (3.452993)	3.332454 (0.8056968)	9 (5.1962)	3 11	0.01123047 0.8921663
3	3 12	0.01650792 0.8716265	11.99713 (6.867643)	10.14216 (5.237033)	4.831156 (1.026541)	12 (6.9282)	3 12	0.006469727 0.8716265
4	3 12	0.0206614 0.8716265	14.96859 (8.501186)	13.11979 (6.888807)	6.036235 (1.184813)	15 (8.6603)	3 13	0.003692627 0.8498418
5	3 12	0.02442433 0.8716265	17.94604 (10.21003)	16.10086 (8.589686)	7.223628 (1.344081)	18 (10.3923)	3 13	0.003692627 0.8498418

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	5 13	0.02452087 0.8603581	9.934442 (4.328975)	4.971259 (1.382582)	4.971259 (1.382582)	10 (4.4721)	5 13	0.02452087 0.8603581
2	5 15	0.01751206 0.8011018	14.96774 (6.608705)	12.78643 (4.676108)	8.783331 (1.906312)	15 (6.7082)	5 15	0.009605408 0.8011018
3	5 15	0.02432663 0.8011018	19.98037 (8.796471)	17.8014 (6.86718)	11.77901 (2.297437)	20 (8.9443)	6 18	0.005311012 0.8275601
4	6 18	0.01745989 0.8275601	29.9498 (12.12675)	27.38701 (9.965438)	17.84703 (2.8659)	30 (12.2474)	6 18	0.005311012 0.8275601
5	6 18	0.02076794 0.8275601	36.00566 (14.59913)	33.43266 (12.42531)	21.36612 (3.261003)	36 (14.6969)	6 19	0.003305376 0.8004705

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	28 45	0.02218546 0.810087	55.81946 (10.34512)	43.76322 (6.122978)	43.76322 (6.122978)	56 (10.583)	28 45	0.02218546 0.810087
2	33 54	0.02116312 0.810509	98.86306 (17.01375)	92.22471 (12.62668)	80.66971 (7.902318)	99 (17.2337)	33 54	0.0114913 0.810509
3	34 56	0.02494717 0.8065126	135.8102 (23.03135)	129.1732 (18.49798)	111.7126 (9.405326)	136 (23.3238)	35 58	0.008010137 0.8026625
4	37 61	0.02400184 0.8158823	184.8859 (30.19847)	177.7495 (25.3255)	152.2413 (11.18461)	185 (30.4138)	38 63	0.006016488 0.8123113
5	38 63	0.02424881 0.8123113	227.6398 (36.6618)	220.49 (31.70998)	188.3215 (12.92202)	228 (36.9865)	40 67	0.004208497 0.8055078

$$\alpha_{ovr} = 0.01, Pointwise\ Power = 0.9$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	4 14	0.006362915 0.9377837	7.983576 (3.957578)	2.744378 (0.740786)	2.744378 (0.740786)	8 (4)	4 14	0.006362915 0.9377837
2	4 15	0.007005218 0.9255533	11.98091 (5.946372)	9.452362 (4.023498)	4.597527 (0.9153506)	12 (6)	4 15	0.003768921 0.9255533
3	4 15	0.009878778 0.9255533	16.00131 (7.936204)	13.46671 (6.015303)	6.166522 (1.056736)	16 (8)	4 16	0.002212524 0.9122111
4	4 16	0.007504341 0.9122111	19.97944 (9.981487)	17.54007 (8.05994)	8.161965 (1.279322)	20 (10)	4 16	0.002212524 0.9122111
5	4 16	0.008991286 0.9122111	23.95095 (11.92227)	21.51719 (10.00686)	9.770237 (1.434854)	24 (12)	5 19	0.001299739 0.9477567

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	8 20	0.009578645 0.9322597	15.97783 (5.59112)	7.869714 (1.738256)	7.869714 (1.738256)	16 (5.6569)	8 20	0.009578645 0.9322597
2	8 22	0.007596185 0.9031455	23.97032 (8.440379)	20.27702 (5.833577)	13.37436 (2.146999)	24 (8.4853)	8 22	0.004065028 0.9031455
3*	9 24	0.009325584 0.9270981	35.97744 (11.93535)	31.71103 (9.116911)	19.6775 (2.510748)	36 (12)	9 25	0.002275692 0.913969
4	9 25	0.007849683 0.913969	45.05927 (14.96858)	40.95815 (12.15911)	25.49438 (2.929551)	45 (15)	9 25	0.002275692 0.913969
5	9 25	0.009439885 0.913969	54.0328 (18.0266)	49.93701 (15.19805)	30.5355 (3.246268)	54 (18)	10 27	0.001966587 0.9347919

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	45 70	0.009405374 0.9006217	89.91218 (13.32989)	69.27616 (7.873896)	69.27616 (7.873896)	90 (13.4164)	45 70	0.009405374 0.9006217
2	52 82	0.008521284 0.9051912	155.9199 (21.58629)	144.5928 (15.53184)	124.2503 (9.572744)	156 (21.6333)	52 82	0.004515245 0.9051912
3	54 86	0.008763323 0.9003806	215.9388 (29.28457)	204.572 (22.97742)	173.9552 (10.99487)	216 (29.3939)	54 86	0.003217405 0.9003806
4	57 91	0.008601433 0.9042405	284.9611 (37.62912)	273.077 (31.05191)	230.017 (12.56655)	285 (37.7492)	57 91	0.002435871 0.9042405
5	58 93	0.008852303 0.9020384	348.1368 (45.76978)	336.224 (39.01762)	281.6142 (13.96219)	348 (45.6946)	60 96	0.001846518 0.9079486

* Design parameters for this row were obtained by substituting 10^3 in the upper limit of the integral for the Type I error formula (see Equation 5.1.2) in the *Des_Inf* code.

$$\alpha_{ovr} = 0.01, Pointwise\ Power = 0.8$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	3 12	0.006469727 0.8716265	5.972426 (3.416178)	2.285659 (0.6955404)	2.285659 (0.6955404)	6 (3.4641)	3 12	0.006469727 0.8716265
2	3 13	0.00684373 0.8498418	8.975906 (5.158271)	7.211382 (3.541614)	3.853152 (0.9260471)	9 (5.1962)	3 13	0.003692627 0.8498418
3	3 13	0.009630278 0.8498418	11.99483 (6.87807)	10.22492 (5.262187)	5.175775 (1.115422)	12 (6.9282)	3 14	0.002090454 0.8269907
4	3 14	0.007054364 0.8269907	14.99365 (8.588582)	13.30582 (6.991712)	6.876503 (1.424767)	15 (8.6603)	3 14	0.002090454 0.8269907
5	3 14	0.008443001 0.8269907	17.99235 (10.35406)	16.3051 (8.744416)	8.231803 (1.645541)	18 (10.3923)	3 15	0.001174927 0.8032494

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	5 15	0.009605408 0.8011018	9.990324 (4.421643)	5.622902 (1.502533)	5.622902 (1.502533)	10 (4.4721)	5 15	0.009605408 0.8011018
2	6 18	0.009842686 0.8275601	17.96538 (7.295522)	15.40048 (5.141215)	10.66882 (2.058423)	18 (7.3485)	6 19	0.003305376 0.8004705
3	6 19	0.008766978 0.8004705	23.99511 (9.752058)	21.58674 (7.628331)	14.93234 (2.624493)	24 (9.798)	6 19	0.003305376 0.8004705
4	7 22	0.006432686 0.8265294	34.98367 (13.16383)	32.19442 (10.82797)	21.82139 (3.197795)	35 (13.2288)	7 22	0.001859583 0.8265294
5	7 22	0.007741562 0.8265294	42.14228 (15.88123)	39.32981 (13.52096)	26.13677 (3.631462)	42 (15.8745)	7 22	0.001859583 0.8265294

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	34 56	0.009593304 0.8065126	67.90258 (11.54017)	54.58797 (6.841786)	54.58797 (6.841786)	68 (11.6619)	34 56	0.009593304 0.8065126
2	39 65	0.009462751 0.808855	116.7911 (18.62893)	109.6118 (13.86542)	97.10712 (8.704557)	117 (18.735)	40 67	0.004208497 0.8055078
3	41 69	0.009540422 0.8022641	163.8427 (25.56137)	156.5856 (20.56941)	137.5359 (10.54688)	164 (25.6125)	43 72	0.003175812 0.8149469
4	44 74	0.009333973 0.811899	219.9841 (33.07177)	212.2223 (27.77891)	184.6081 (12.41654)	220 (33.1662)	45 76	0.002227512 0.8089351
5	45 76	0.009522206 0.8089351	269.9505 (40.11444)	262.1602 (34.72178)	227.129 (14.2406)	270 (40.2492)	46 78	0.001865723 0.8060517

$$\alpha_{ovr} = 0.001, Pointwise\ Power = 0.9$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	5 20	0.0007719398 0.9387666	10.0123 (4.485938)	3.9329 (0.8907877)	3.9329 (0.8907877)	10 (4.4721)	5 20	0.0007719398 0.9387666
2	5 21	0.0008743739 0.9289592	15.01158 (6.712832)	12.06768 (4.55437)	6.415964 (1.085284)	15 (6.7082)	5 21	0.0004552603 0.9289592
3	5 22	0.0007487147 0.9183452	20.02216 (8.963022)	17.16854 (6.807211)	8.975901 (1.307772)	20 (8.9443)	5 22	0.0002667606 0.9183452
4	5 22	0.0009706419 0.9183452	24.98507 (11.14538)	22.14079 (9.000686)	11.20995 (1.496245)	25 (11.1803)	5 23	0.0001553744 0.9069417
5	5 23	0.000698678 0.9069417	30.07172 (13.47475)	27.30345 (11.32346)	13.97195 (1.77373)	30 (13.4164)	5 23	0.0001553744 0.9069417

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni true Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	10 29	0.0008290263 0.910977	19.98268 (6.341044)	11.39437 (2.099073)	11.39437 (2.099073)	20 (6.3246)	10 29	0.0008290263 0.910977
2†	11 32	0.0009081847 0.9209501	32.97134 (9.96132)	28.28809 (6.872559)	19.50176 (2.562185)	33 (9.9499)	11 32	0.000470337 0.9209501
3†	11 33	0.0008571444 0.9090848	44.00483 (13.25985)	39.49741 (10.18349)	26.837 (3.018996)	44 (13.2665)	11 33	0.000303053 0.9090848
4	12 35	0.0009886811 0.9296741	59.92714 (17.282)	54.85737 (14.01526)	35.72402 (3.378549)	60 (17.3205)	12 36	0.0001730006 0.9192982
5	12 36	0.000789422 0.9192982	72.04075 (20.83322)	67.13102 (17.5646)	43.89268 (3.880026)	72 (20.7846)	12 36	0.0001730006 0.9192982

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected Person Years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	66 107	0.0008453577 0.9057528	131.9755 (16.2538)	106.1604 (9.778866)	106.1604 (9.778866)	132 (16.2481)	66 107	0.0008453577 0.9057528
2	69 113	0.0009906176 0.9001822	207.0493 (24.9882)	194.0932 (18.03552)	170.8481 (11.23494)	207 (24.9199)	71 116	0.00045905 0.9059773
3	73 120	0.0009389094 0.9025216	291.8217 (34.19647)	278.4314 (26.84334)	242.3819 (13.03819)	292 (34.176)	73 120	0.0003275152 0.9025216
4	76 125	0.000940831 0.9064813	379.9758 (43.43582)	366.0709 (35.84505)	315.5337 (14.68223)	380 (43.589)	76 125	0.0002497132 0.9064813
5	77 127	0.000980573 0.9048674	462.3637 (52.70097)	448.3553 (44.90322)	384.1871 (16.23976)	462 (52.6498)	78 129	0.0001781977 0.9032685

† Bonferroni values for this row were obtained by substituting 10^3 in the upper limit of the integral for the Type I error formula in the *Des_Inf* code.

$$\alpha_{ovr} = 0.001, Pointwise\ Power = 0.8$$

$r = 10$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	4 18	0.000744819 0.8824026	7.995816 (3.995293)	3.467864 (0.8617723)	3.467864 (0.8617723)	8 (4)	4 18	0.000744819 0.8824026
2	4 19	0.0008204864 0.8660641	11.99013 (5.988192)	9.817389 (4.118107)	5.674194 (1.12055)	12 (6)	4 19	0.0004277229 0.8660641
3	4 20	0.0006840587 0.8488657	16.00544 (8.001619)	13.91007 (6.139492)	7.946991 (1.431762)	16 (8)	4 20	0.0002441406 0.8488657
4	4 20	0.0008862584 0.8488657	19.97916 (9.97136)	17.88923 (8.113325)	9.921867 (1.689673)	20 (10)	4 20	0.0002441406 0.8488657
5	4 21	0.0006217585 0.8308855	24.0397 (12.04926)	22.02267 (10.19584)	12.37588 (2.08553)	24 (12)	4 21	0.0001385808 0.8308855

$r = 5$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	8 26	0.0006593636 0.827317	16.00334 (5.674045)	9.971721 (2.007542)	9.971721 (2.007542)	16 (5.6569)	8 26	0.0006593636 0.827317
2	8 27	0.0007923881 0.8050986	23.98932 (8.492664)	21.14072 (6.069242)	15.9089 (2.614148)	24 (8.4853)	8 27	0.0004106977 0.8050986
3	9 30	0.0006685265 0.8292057	35.96352 (11.98107)	32.75027 (9.356992)	23.85317 (3.243882)	36 (12)	9 30	0.0002359934 0.8292057
4	9 30	0.000869796 0.8292057	45.06092 (15.00252)	41.82748 (12.35957)	29.80425 (3.772347)	45 (15)	9 30	0.0002359934 0.8292057
5	9 31	0.0006700265 0.8087628	54.06693 (18.1057)	50.99137 (15.4765)	36.63959 (4.590356)	54 (18)	9 31	0.0001470384 0.8087628

$r = 2$								
Number of experimental treatment groups	Number of control group events, critical value	True Type I error, true pointwise power	Expected person years under null (std dev)	Expected person years under specified alternatives (std dev)		Expected person years in an uncurtailed study (std dev)	Bonferroni control group events, critical value	Bonferroni True Type I error, true power
				One Tx group meets the rate	All Tx groups meet the rate			
1	51 87	0.0009943432 0.8101087	101.9314 (14.28051)	85.28772 (8.573397)	85.28772 (8.573397)	102 (14.2829)	51 87	0.0009943432 0.8101087
2	55 95	0.000956185 0.800315	164.9128 (22.24269)	156.5008 (16.64607)	141.8629 (10.54808)	165 (22.2486)	55 95	0.0004916666 0.800315
3	59 102	0.0009015878 0.8075131	236.0527 (30.75402)	227.0928 (24.70533)	203.6122 (12.73727)	236 (30.7246)	59 102	0.0003146735 0.8075131
4	60 104	0.0009923353 0.8053415	299.9331 (38.7307)	290.972 (32.62529)	259.2723 (14.93282)	300 (38.7298)	61 106	0.0002216631 0.8032139
5	62 108	0.0008655343 0.8011291	371.8901 (47.327)	362.8875 (41.04687)	322.5611 (17.22202)	372 (47.244)	62 108	0.0001860506 0.8011291

We next provide the formulas for partial power and full power in the context of testing for treatment inferiority. Partial power denotes the probability that at least one truly inferior experimental treatment is found to be inferior to the control, assuming that the alternative hypothesis in Equation 5.1.1 is true. To calculate partial power, we compute $P(\max(D_1, \dots, D_K) \geq w)$ under the alternative hypothesis in Equation 5.1.1. The result again follows from Equation 4.2.5 and is given by

$$P(\max(D_1, \dots, D_K) \geq w) =$$
$$1 - \int_0^\infty \frac{t^{d_C-1}e^{-t}}{\Gamma(d_C)} \left[\sum_{s=0}^{w-1} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right] \left[\sum_{s=0}^{w-1} e^{-r_2 t} \frac{(r_2 t)^s}{s!}\right] \dots \left[\sum_{s=0}^{w-1} e^{-r_K t} \frac{(r_K t)^s}{s!}\right] dt \quad (5.1.4)$$

Full power corresponds to the probability that all truly inferior experimental treatments are found to be inferior to the control, assuming that the alternative hypothesis in Equation 5.1.1 is true. In this case, the number of events in all truly inferior experimental treatment groups must be sufficiently large. Hence, we must find $P(\min(D_1, \dots, D_K) \geq w)$ under the alternative hypothesis in Equation 5.1.1 and we proceed as follows:

$$P(\min(D_1, \dots, D_K) \geq w|t) = 1 - P(\min(D_1, \dots, D_K) < w|t)$$
$$= 1 - P(\min(D_1, \dots, D_K) \leq w - 1|t)$$

Therefore, by Equation 4.2.4, we have

$$P(\min(D_1, \dots, D_K) \geq w) =$$
$$\int_0^\infty \frac{t^{d_C-1}e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{w-1} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right] \dots \left[1 - \sum_{s=0}^{w-1} e^{-r_K t} \frac{(r_K t)^s}{s!}\right] dt \quad (5.1.5)$$

As was noted in Section 4, the number of elements in Equations 5.1.4 and 5.1.5 should, in theory, be appropriately reduced to the number of truly inferior experimental treatments, though this value will not be known in practice.

Recall that an important consideration for investigators is the expected number of person years of follow-up until trial termination, as the longer a study lasts or the larger the number of subjects which need to be enrolled, the greater the expenses. We therefore turn our attention to calculating the expected number of person years until trial termination in an inferiority study. We will consider trials conducted under both uncurtailed and fully curtailed stoppage.

In an uncurtailed design, the trial will stop only when the control group reaches d_C events, so the formulas for the expected number and variance of person years are given by $(K+1) \times \frac{d_C}{i_C}$ and $(K+1)^2 \times \frac{d_C}{{i_C}^2}$, respectively (the same as in an uncurtailed superiority trial). We will, as in Section 4.6, assume that $i_C = 1$ for the following derivations concerning subject follow-up in inferiority trials, keeping in mind that, in practice, multiplication by $\frac{1}{i_C}$ will need to be performed when reporting the expected amount and standard deviation of follow-up.

In a fully curtailed inferiority trial, the stopping rules are as follows:

- Follow-up is stopped for any experimental treatment arm that reaches w events prior to the control arm reaching d_C events and H_0 is rejected for all such experimental treatment arms.
- The entire study is stopped if the control arm reaches d_C events (before all experimental treatment arms reach w events), and H_0 is accepted for all experimental treatment arms which have not yet reached w events.
- If all experimental treatment arms reach w events before the control arm reaches d_C events, then the entire study is stopped and H_0 is rejected for all experimental treatment arms.

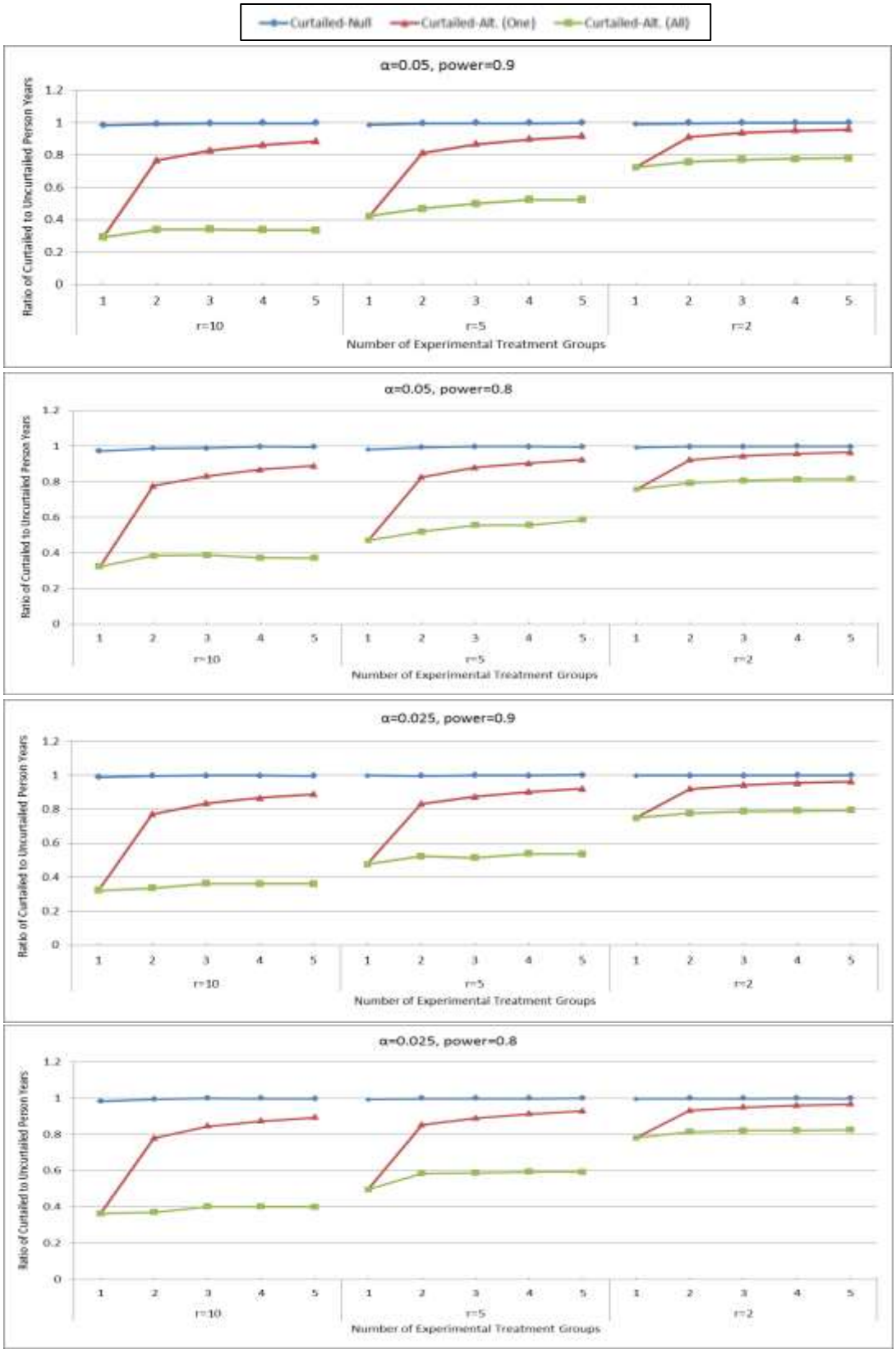
The algorithms to calculate the expected number and standard deviation of person years of follow-up until trial termination in a fully curtailed inferiority study are identical to those presented in Section 4.6 for superiority trials, with one modification. In Section 4.6, follow-up of an experimental treatment arm was discontinued once it reached $m + 1$ events. In the inferiority setting, follow-up of an experimental treatment arm will terminate once it reaches w events. Hence, replacing $m + 1$ with w in the algorithms in Section 4.6 will yield the desired results for inferiority studies. Estimates of the expected number and standard deviation of person years of follow-up until study termination for inferiority trials are included in Table 4 for the specified parameter combinations[39]. Values are computed under the null hypothesis (see column 4) and under two specific alternative hypotheses: (1) one new treatment group has the rate ratio r specified in the table, and the remaining $K - 1$ new treatment groups have rate ratio equal to one (see column 5), and (2) the indicated value of r in the table holds for all new treatments under study (i.e. $r_1 = r_2 = \cdots = r_K = r$) (see column 6). Table 4 also contains the expected number and standard deviation of person years until trial discontinuation for an uncurtailed study (with $i_C = 1$) in column 7.

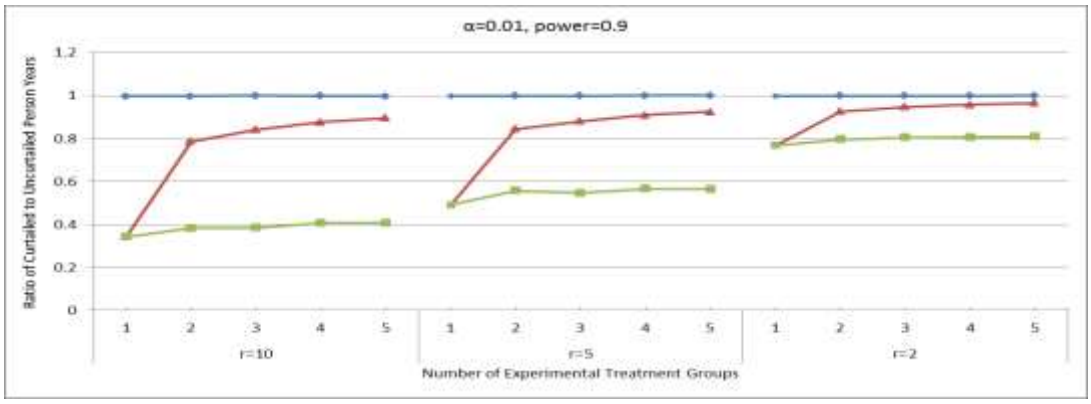
Figure 6 displays the ratio of curtailed to uncurtailed expected person years in an inferiority trial for each combination of design parameters in Table 4; in particular, the ratios of the values in column 4 to column 7 (labeled "Curtailed-Null"), column 5 to column 7 (labeled "Curtailed-Alt (One)"), and column 6 to column 7 (labeled "Curtailed-Alt (All)") in Table 4 are plotted in Figure 6. Figure 6 illustrates that the expected number of person years is reduced the most in a fully curtailed inferiority trial when the

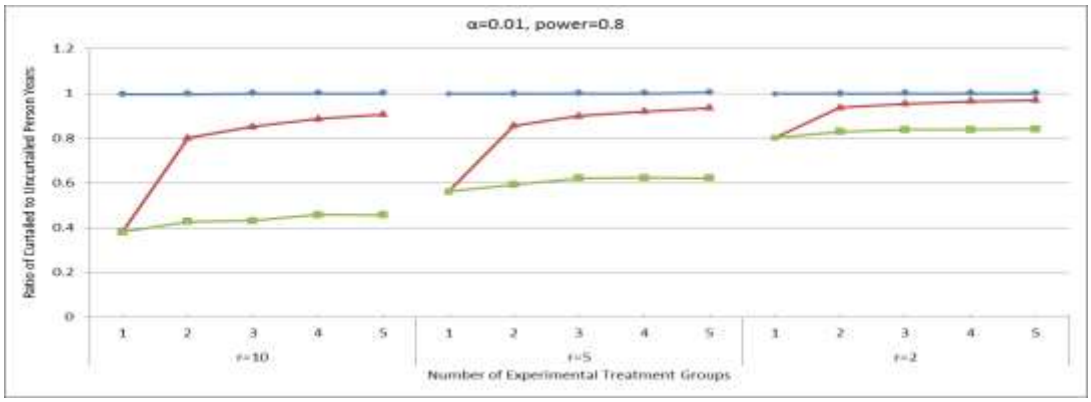
[39] The person year values in columns 4 through 6 of Table 4 were estimated using *Null_Time* and *Alt_Time* with 100,000 simulations; note that these functions return results based on assuming $i_C = 1$.

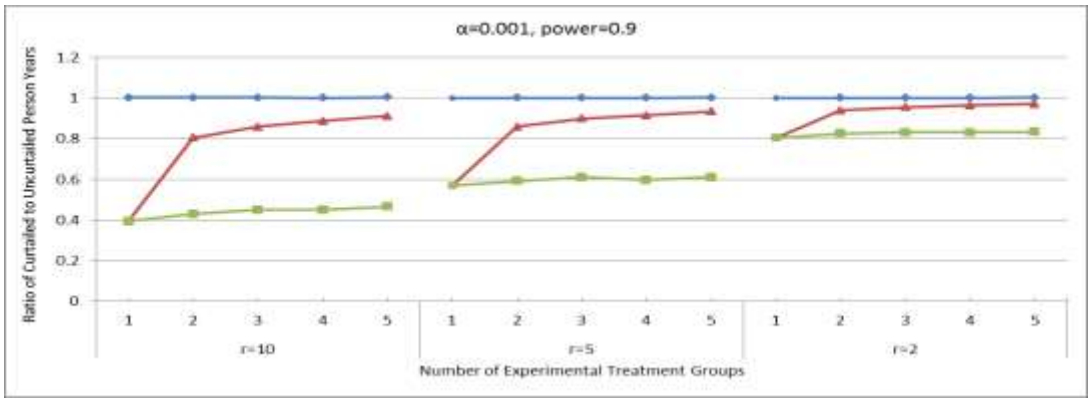
alternative hypothesis is true. This is because events tend to accumulate more quickly in the experimental treatment arms under the alternative hypothesis than the null hypothesis, and hence reach *w* events (the time at which follow-up of the experimental treatment arms can be terminated) more quickly.

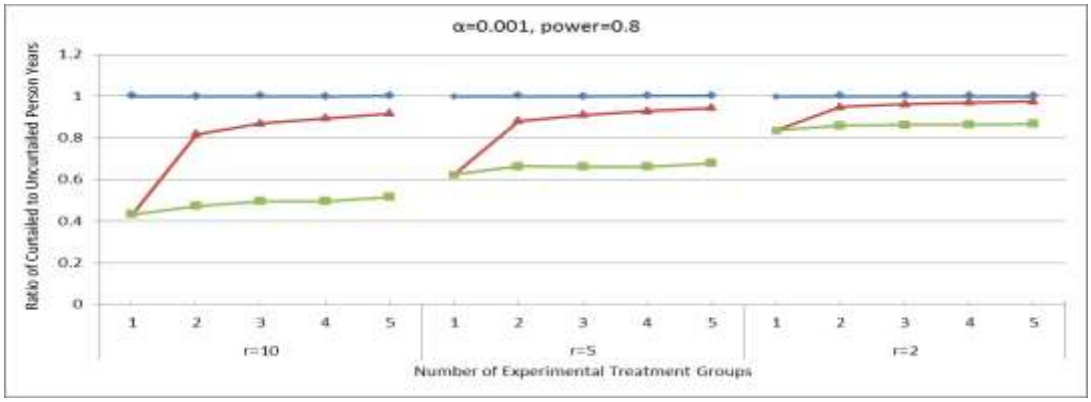
Figure 6: Ratio of fully curtailed to uncurtailed expected person years of follow-up for inferiority trials with design parameters given in Table 4




α=0.01, power=0.9
Ratio of Curtailed to Uncurtailed Person Years
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
r=10
r=5
r=2
Number of Experimental Treatment Groups


α=0.01, power=0.8
Ratio of Curtailed to Uncurtailed Person Years
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
r=10
r=5
r=2
Number of Experimental Treatment Groups


α=0.001, power=0.9
Ratio of Curtailed to Uncurtailed Person Years
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
r=10
r=5
r=2
Number of Experimental Treatment Groups


α=0.001, power=0.8
Ratio of Curtailed to Uncurtailed Person Years
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
r=10
r=5
r=2
Number of Experimental Treatment Groups

5.2: Two-sided Test Combining One-sided Superiority and Inferiority Boundaries

We now present a two-sided test which combines the information obtained from the test of the hypotheses in Equation 4.2.1 corresponding to demonstrating treatment superiority and the test of the hypotheses in Equation 5.1.1 corresponding to demonstrating treatment inferiority. The proposed two-sided test will allow researchers to determine whether rejection of the null hypothesis for an individual experimental treatment group is due to superiority of the experimental treatment in comparison to the control (i.e. a sufficiently small number of events occur in the experimental treatment group compared to the control) or is due to inferiority of the experimental treatment in comparison to the control (i.e. too many events occur in the experimental treatment group compared to the control)[40].

Based on these considerations, the two sided hypotheses are as follows:

$$H_0: i_1 = i_2 = \cdots = i_K = i_C \; versus \; H_a: i_1 = r_1 i_C, i_2 = r_2 i_C, \ldots, i_K = r_K i_C \quad \text{where at least one of the } r_k \neq 1 \qquad (5.2.1)$$

We will reject the null hypothesis that all experimental treatments are statistically equivalent to the control in terms of efficacy in favor of the alternative hypothesis when either few enough or too many events occur in at least one of the experimental treatment groups.

[40] Throughout Section 5.2, when a sufficiently small number of events are observed in a new treatment group we declare the new treatment to be superior to the control, as we have implicitly taken "events" to mean number of disease cases. This may be an appropriate assumption in the context of most clinical trials, but in other, perhaps non-clinical, settings, rejection due to too few or too many events may both be indicative of an undesirable comparator. Consider the following example: suppose there are several new radiation-detection devices which need to be calibrated. Each will be exposed to the same number of radioactive particles, as controlled by study investigators, and compared to a control device which is known to detect the amount of radiation with sufficient accuracy. If a new device reads too few or too many particles, it will be declared inadequately calibrated. In either case, misreading the amount of radiation could have significant practical consequences, and, as such, rejection in either direction is indicative of an inadequate comparator.

We can construct a conservative two-sided test with a two-sided overall Type I error of α_{ovr} for the hypotheses in Equation 5.2.1 by combining the one-sided tests for superiority (see Section 4.2) and inferiority (see Section 5.1) using the Bonferroni approach. We define a "balanced two-sided test" to be one in which the two-sided overall Type I error is equally allocated to the one-sided superiority and inferiority tests; that is, the parameters for the balanced two-sided test are derived from the one-sided superiority and inferiority tests each at significance level $\alpha_{ovr}/2$. By the Bonferroni method, the two-sided overall Type I error of the balanced two-sided test is $\leq \alpha_{ovr}/2 + \alpha_{ovr}/2 = \alpha_{ovr}$. When in addition the superiority test is powered using rate ratio r and the inferiority test is powered using rate ratio $1/r$ (power is taken to be pointwise for both of the one-sided tests), the test will be referred to as a "completely balanced two-sided test". In contrast, the term "unbalanced two-sided test" will denote tests for which the level of significance allocated to the one-sided superiority and inferiority tests are α_1 and α_2, respectively, where $\alpha_1 + \alpha_2 = \alpha_{ovr}$ and $\alpha_1 \neq \alpha_2$.

To construct the two-sided test at overall significance level $\alpha_1 + \alpha_2 = \alpha_{ovr}$, we need to find the following parameters:

1. The number of events to observe in the control group d_C when testing for superiority at significance level α_1. For the remainder of this subsection, we will refer to this value as d_{CS}, the subscript *S* identifying that the value is obtained from a superiority trial.
2. The critical value m from the test of superiority at significance level α_1.
3. The number of events to observe in the control group d_C when testing for inferiority at significance level α_2. For the remainder of this subsection, we will refer to this value as d_{CI}, the subscript *I* identifying that the value is obtained from an inferiority trial.

4. The critical value w from the test of inferiority at significance level α_2.

Hence, there are four values (d_{CS}, m, d_{CI}, and w) needed to construct the two-sided test. There are six possible arrangements[41] for these values, each of which is displayed on a number line in Figure 7 below, where the x-axis is discrete and enumerates number of events.

For a balanced two-sided test we require that $\alpha_1 = \alpha_2 = \alpha_{ovr}/2$, and for a completely balanced two-sided test we impose the additional restriction that the parameters d_{CS} and m are determined from a superiority trial powered using rate ratio r and the parameters d_{CI} and w are determined from an inferiority trial powered using rate ratio $1/r$.

[41] For the sake of brevity, in Figure 7 we do not present cases in which the parameter values coincide. Based on the parameter values in Tables 2 and 4, of the 120 completely balanced two-sided tests we observed 12 cases where $m = d_{CI}$, 94 cases where $w = d_{CS}$, and 14 cases where there was no overlap of parameter values. No test had more than two coinciding parameter values. Also for the completely balanced two-sided tests constructed from the values in Tables 2 and 4, we observed only configurations (1), (2), and (3) in Figure 7. However, we suspect that it may be possible to obtain the remaining configurations in the case of the more general unbalanced two-sided test.

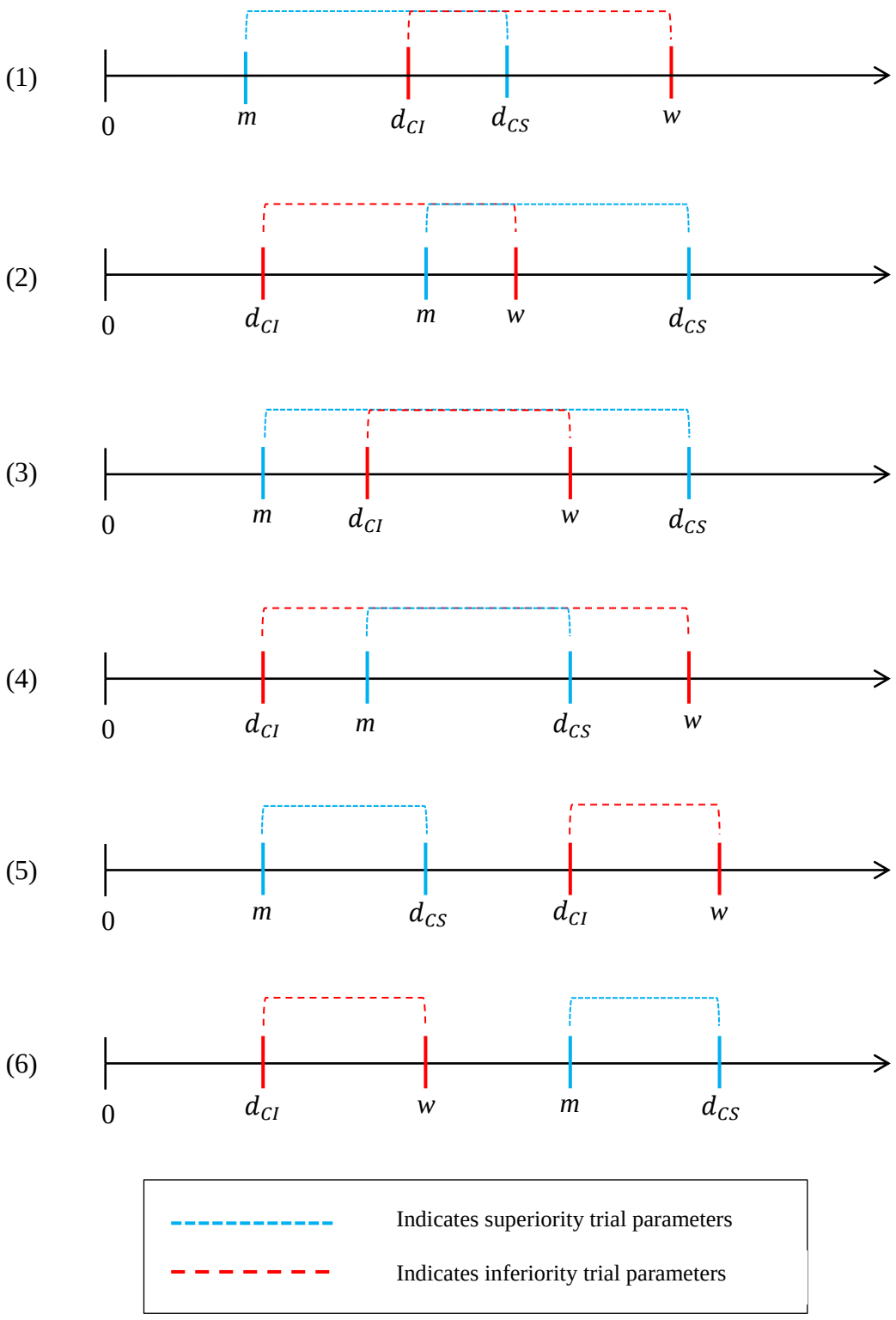


Figure 7: Possible arrangements of d_{CS}, m, d_{CI}, and w in a two-sided test

Note that it is impractical to conduct an uncurtailed trial in the two-sided test setting. This is because if the study were to continue until the control group reaches $\max(d_{CS}, d_{CI})$, then some of the new treatment groups would have needless (and hence uneconomical) follow-up depending on the configuration of the study parameters, or the decision made at $\min(d_{CS}, d_{CI})$ for each new treatment group could be contradicted at $\max(d_{CS}, d_{CI})$. Consider configurations (1) through (4) in Figure 7; in these configurations, $\max(d_{CS}, d_{CI}) = d_{CS}$ and $m < w$. These configurations imply that the control group reaches d_{CI} events before d_{CS} events, at which time the new treatments would be evaluated for inferiority. If any of the new treatments have reached w events at this time, they would be declared inferior to the control, and hence it would not make statistical nor economic sense to continue follow-up in these arms and subsequently evaluate them for superiority once the control group reaches d_{CS} events, as they have also already exceeded m events (since $w > m$). Now, consider configuration (5) in Figure 7 in which the control group reaches d_{CS} events prior to d_{CI} events. Suppose a given new treatment group has not reached $m + 1$ events by the time the control group reaches d_{CS} events. The new treatment will then be declared to be superior to the control. If the trial is uncurtailed, the new treatment group will continue to be followed until the control group reaches d_{CI} events. As the trial continues, if the new treatment group accumulates w events prior to the control group reaching d_{CI} events, it would indicate that the new treatment is inferior to the control, thus contradicting the decision made when superiority was evaluated at the time the control group reached d_{CS} events. Similar rationale applies to configuration (6). These examples illustrate that it is not sensible to conduct an uncurtailed trial in the two-sided test setting.

A fully curtailed trial is possible for the two-sided setting, and the rules of stoppage for each treatment arm are as follows:

- If a given new treatment arm reaches w events prior to the control group reaching d_{CI} events, then discontinue follow-up in the new treatment arm and declare the new treatment to be inferior to the control.
- If at the time the control group reaches d_{CI} events a given new treatment arm has accrued $[m+1, w-1]$ events, discontinue follow-up in the new treatment arm and declare the new treatment to be neither inferior nor superior to the control.
- While the number of events observed in the control group is in $[d_{CI}+1, d_{CS}]$, if a given new treatment arm reaches $m+1$ events, discontinue follow-up in the new treatment arm and declare the new treatment to be neither inferior nor superior to the control.
- If a given new treatment arm has accrued less than $m+1$ events at the time the control group reaches d_{CS} events, then discontinue follow-up in the new treatment arm and declare the new treatment to be superior to the control.
- Follow-up of the control arm is terminated at the earliest of (1) a decision is made for all new treatment arms (i.e. there are no longer any active new treatment arms because each new treatment has been declared inferior, superior, or neither inferior nor superior to the control) and (2) the control arm reaches $\max(d_{CS}, d_{CI})$ events.

Though the rules above apply, in theory, to configurations (5) and (6), their practical application to these configurations is questionable. This is because in configuration (5), the superiority boundaries occur prior to the inferiority boundaries, and vice versa for

configuration (6). In configuration (5), follow-up of all new treatment arms which have $\leq m$ events when the control arm reaches d_{CS} events will be terminated and the new treatments declared superior to the control, while all new treatment arms with $\geq m + 1$ events are followed until the control arm reaches d_{CI} events, at which time these treatments are evaluated for inferiority. These rules indicate that the new treatment arms that are declared superior to the control at the time the control reaches d_{CS} events are not evaluated for inferiority. It could be argued that if these treatment arms were followed until the control arm reaches d_{CI} events, they may accumulate w events, leading to a seemingly contradictory classification of the treatments as inferior. Take for example a trial with parameters $m = 3, d_{CS} = 9, d_{CI} = 12$, and $w = 17$, and suppose that one of the new treatments is declared superior to the control as the treatment group has accrued ≤ 3 events by the time the control arm reaches $d_{CS} = 9$ events. Rather than terminate follow-up in this new treatment arm per the curtailment rules, suppose instead that follow-up of this new treatment arm were continued until the control arm reaches d_{CI} events. If at this time the number of events in the new treatment arm is ≥ 17, then the treatment would be declared inferior to the control as it has reached the value of w. Researchers may therefore be concerned that configuration (5) could lead to an improper designation of treatment superiority as not all new treatments are evaluated for both superiority and inferiority. Similar reasoning also applies to configuration (6). However, such a reversal is unlikely to happen (i.e. by chance alone), unless a time-mediated change in the rate of events is observed.

We next provide an example of the completely balanced two-sided test. Suppose there are five experimental treatments under study and we wish to determine whether any

are statistically significantly different from the control at a two-sided overall Type I error $\alpha_{ovr} = 0.05$, power equal to 0.8, and for a rate ratio of $r = 0.2$ when an experimental treatment is superior to the control and a rate ratio of $1/r = 1/0.2 = 5$ when the experimental treatment is inferior to the control. We use *Des_Sup(5,0.025,0.2,0.9)* and *Des_Inf(5,0.025,5,0.9)* to find the trial design parameters (i.e. we find the design parameters using the one-sided trial parameters at level of significance $\frac{a_{ovr}}{2} = 0.025$ and at pointwise power $\sqrt{.8} \approx 0.9$)[42]. Using these functions (or extracting their values from Tables 2 and 4), we find the design parameters $d_{CS} = 22$, $m = 7$, $d_{CI} = 8$, and $w = 21$. These parameters are consistent with configuration (3) in Figure 7.

5.3: Two-sided Test Based on the Chi-squared Distribution for Detection of a Cumulative Signal

The hypotheses in Equation 5.2.1 can also be tested using a test-statistic which follows a chi-squared distribution under the null hypothesis. This test will reject based on the cumulative difference of the experimental treatment groups from the control; that is, the probability of rejection increases when an experimental treatment is either superior or inferior to the control treatment.

Steyn (1955) proved that for $X_1, X_2, \dots, X_n \sim NM(\nu, p_0, p_1, \dots, p_n)$, the statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(x_i - \nu \frac{p_i}{p_0}\right)^2}{\nu \frac{p_i}{p_0}} - \frac{\left(\nu + \sum_{i=1}^{n} x_i - \frac{\nu}{p_0}\right)^2}{\frac{\nu}{p_0}}$$

asymptotically follows a chi-squared distribution with n degrees of freedom. Since $D_1, D_2, \dots, D_K | d_C \sim NM\left(d_C, \frac{i_C}{i_C + \sum_{k=1}^{K} i_k}, \frac{i_1}{i_C + \sum_{k=1}^{K} i_k}, \frac{i_2}{i_C + \sum_{k=1}^{K} i_k}, \dots, \frac{i_K}{i_C + \sum_{k=1}^{K} i_k}\right)$ under Design C, we know that

[42] See Section 6 for a discussion of the power of the two-sided test.

$$\chi^2 = \sum_{j=1}^{K} \frac{\left(d_j - d_C \dfrac{\dfrac{i_j}{i_C + \sum_{k=1}^{K} i_k}}{\dfrac{i_C}{i_C + \sum_{k=1}^{K} i_k}} \right)^2}{d_C \dfrac{\dfrac{i_j}{i_C + \sum_{k=1}^{K} i_k}}{\dfrac{i_C}{i_C + \sum_{k=1}^{K} i_k}}} - \frac{\left(d_C + \sum_{j=1}^{K} d_j - \dfrac{d_C}{\dfrac{i_C}{i_C + \sum_{k=1}^{K} i_k}} \right)^2}{\dfrac{d_C}{\dfrac{i_C}{i_C + \sum_{k=1}^{K} i_k}}}$$

asymptotically follows a chi-squared distribution with K degrees of freedom, where d_j is an observed value of the random variable D_j for $j = 1,2, \ldots, K$. Under the null hypothesis in Equation 5.2.1, the negative multinomial distribution is balanced, and so $i_1 = i_2 = \cdots = i_K = i_C$. Thus, under the null hypothesis, the chi-squared test statistic is

$$\chi^2_{null} = \sum_{j=1}^{K} \frac{\left(d_j - d_C\right)^2}{d_C} - \frac{\left(d_C + \sum_{j=1}^{K} d_j - \dfrac{d_C}{\dfrac{1}{K+1}} \right)^2}{\dfrac{d_C}{\dfrac{1}{K+1}}}$$

$$= \sum_{j=1}^{K} \frac{\left(d_j - d_C\right)^2}{d_C} - \frac{\left(\sum_{j=1}^{K} d_j - d_C K\right)^2}{d_C(K+1)}$$

Thus, we will reject the null hypothesis in Equation 5.2.1 in favor of the alternative at level of significance α when χ^2_{null} exceeds the critical value $\chi^2_{\alpha,K}$, where $P\left(\chi^2_K > \chi^2_{\alpha,K}\right) = \alpha$.

To illustrate this test, consider an agricultural experiment in which five plants of the same species are each exposed to a different agent, four being experimental pesticides and one being a control treatment (no exposure to pesticides). The outcome variable in this experiment is the number of holes in each plant due to insect activity. Assuming holes appear in the leaves over time according to a Poisson process, the NMD characterization is appropriate for this study given that we stop the experiment once a

pre-specified number of holes appear in the control plant. We will use the chi-squared test to determine if the ability of any of the pesticides in repelling insects is different from that of the control treatment. Suppose we wait until $d_C = 20$ holes are found in the leaves of the control plant, and, at that time, the number of holes found in the four remaining plants are $d_1 = 16, d_2 = 13, d_3 = 23$, and $d_4 = 18$. Then, our test statistic is

$$\chi^2_{null} = \sum_{j=1}^{4} \frac{\left(d_j - 20\right)^2}{20} - \frac{(70 - (20 * 4))^2}{(20 * 5)} = 0.8 + 2.45 + 0.45 + 0.2 - 1 = 2.9$$

If we conduct the test at significance level 0.05, the critical values is $\chi^2_{0.05,4} \approx 9.488$. Hence, since our test statistic does not exceed the critical value at 5% significance, we do not have enough evidence to conclude that the efficacy of any of the experimental pesticides in protecting the plant species from insect damage is different from that of the control treatment.

As a final example of the application of the chi-squared test, suppose a manufacturer is evaluating five potential suppliers to provide a machine component. The manufacturer wants to know if any of the potential suppliers produce the component with a different rate of breakdown than the current supplier, so that the manufacturer can evaluate prospective business partners (we assume the number of breakdowns over time follows a Poisson distribution so that the chi-squared test is applicable). As such, the manufacturer orders one component from each of the potential suppliers and records the number of times each one breaks down. The experiment is terminated when the current supplier's component reaches 12 breakdowns. Suppose that the number of breakdowns at this time among the five potential suppliers' machine components is $d_1 = 7, d_2 = 5, d_3 = 16, d_4 = 3$, and $d_5 = 3$. The test statistic is thus

$$\chi^2_{null} = \sum_{j=1}^{5} \frac{\left(d_j - 12\right)^2}{12} - \frac{\left(34 - (12 * 5)\right)^2}{(12 * 6)} = \frac{25}{12} + \frac{49}{12} + \frac{4}{3} + \frac{27}{4} + \frac{27}{4} - \frac{169}{18} \approx 11.611$$

The critical value at 5% significance is approximately 11.071. Thus, at 5% significance, the null hypothesis is rejected and the manufacturer can conclude that at least one potential supplier provides a machine component with a different rate of breakdown than the component of their current supplier. One limitation to this test, however, is that it does not identify which supplier(s) are different, albeit the most extreme deviations could perhaps be identified qualitatively. The manufacturer would want to subsequently identify which supplier(s) can provide a component less prone to failures and may also want to know which supplier(s) produce machine components which are prone to high rates of failure so that they can avoid using these suppliers in the future.

This lack of identification of which new condition(s) is different from the control is an important difference between the chi-squared test and the two-sided test combining the one-sided superiority and inferiority boundaries discussed in Section 5.2. In the context of clinical studies, the two-sided test based on the chi-squared distribution rejects based on the cumulative difference of the experimental treatment groups from the control. That is, the value of the test statistic increases when a treatment is either superior or inferior to the control. However, if the null hypothesis in Equation 5.2.1 is rejected, the chi-squared test does not indicate which of the experimental treatment groups are responsible for the rejection (i.e. it does not indicate which experimental treatments are superior or inferior to the control). This information is often desired by researchers, making the usefulness of the chi-squared test limited. In contrast, the two-sided test which combines the one-sided superiority and inferiority boundaries does identify which

individual experimental treatments are superior or inferior to the control and responsible for rejection of the null hypothesis in Equation 5.2.1.

5.4: Pairwise Tests of Experimental Treatments

Thus far, all evaluations of the superiority or inferiority of experimental treatments have been conducted in comparison to a control treatment. Another interest may be in comparing the efficacy of the experimental treatments to each other, even if the study is being conducted as a negative multinomial process with stoppage at a specified number of events in the control group. Consider if multiple experimental treatments are declared superior to the control, then investigators will likely want to further know which experimental treatment is the most effective so that resources can be invested into this treatment. This is important as the expenses associated with development of a drug increase as the drug advances through the necessary clinical trials for approval.

Dose ranging studies provide a natural setting for comparing experimental treatments to each other. Though such studies may include a control group, the primary objective is to compare varying doses of a drug to determine the minimal effective and maximum tolerable doses so that an optimal dosing strategy can be determined for subsequent clinical trials.

In this dissertation, comparison of the experimental treatments to each other will depend upon the fact that if $X_1, X_2, \dots, X_n \sim NM(\nu, p_0, p_1, p_2, \dots, p_n)$, then conditional on the sum $X_1 + X_2 + \cdots + X_n$, the distribution of $X_1, X_2, \dots, X_n$ is multinomial with parameters $\sum_{i=1}^{n} x_i$ and $\frac{p_j}{\sum_{i=1}^{n} p_i}, j = 1, \dots, n$. The proof is as follows[43]:

[43] The desired result is stated (without a formal proof) in Lemma 1 in Tsui (1986, pages 47-48). Lemma 1 also states that the distribution of $\sum_{i=1}^{n} X_i$ is negative binomial with parameters ν and $1 - \sum_{i=1}^{n} p_i$, a fact that is used in our proof.

$$P\left(X_1 = x_1, \dots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = z\right) = \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(\sum_{i=1}^{n} X_i = z)}$$

$$= \frac{\frac{\Gamma(\nu + \sum_{i=1}^{n} x_i)}{\Gamma(\nu)\prod_{i=1}^{n} x_i!}(1 - \sum_{i=1}^{n} p_i)^{\nu}\prod_{i=1}^{n} p_i^{x_i}}{\frac{\Gamma(\nu + \sum_{i=1}^{n} x_i)}{\Gamma(\nu)(\sum_{i=1}^{n} x_i)!}(1 - \sum_{i=1}^{n} p_i)^{\nu}(\sum_{i=1}^{n} p_i)^{\sum_{i=1}^{n} x_i}} = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!}\frac{\prod_{i=1}^{n} p_i^{x_i}}{(\sum_{i=1}^{n} p_i)^{x_1 + x_2 + \cdots + x_n}}$$

$$= \binom{\sum_{i=1}^{n} x_i}{x_1, x_2, \dots, x_n}\left(\frac{p_1}{\sum_{i=1}^{n} p_i}\right)^{x_1}\left(\frac{p_2}{\sum_{i=1}^{n} p_i}\right)^{x_2}\cdots\left(\frac{p_n}{\sum_{i=1}^{n} p_i}\right)^{x_n}$$

$$\sim Multinomial\left(\sum_{i=1}^{n} x_i, \frac{p_1}{\sum_{i=1}^{n} p_i}, \frac{p_2}{\sum_{i=1}^{n} p_i}, \dots, \frac{p_n}{\sum_{i=1}^{n} p_i}\right)$$

So, pairwise comparisons of experimental treatment groups can be conducted using the multinomial distribution. Under Design C, by Equation 2.4.1 we know that the parameters in our multinomial distribution are $\sum_{k=1}^{K} d_k$ and $i_j / \sum_{k=1}^{K} i_k$ for $j = 1,2, \dots, K$.

An exact multiple comparisons test for the multinomial distribution was developed by Shaffer (1971). For a multinomial distribution with k outcome categories, Shaffer's exact test simultaneously tests the $k(k-1)$ hypotheses of the form $H_0: p_i \leq p_j \; vs \; H_a: p_i > p_j$ and $H_0: p_i \geq p_j \; vs \; H_a: p_i < p_j$ for all $i \neq j$. For example, for a trinomial distribution there are six hypotheses as listed below:

$$H_0^1: p_1 \leq p_2 \; vs \; H_a^1: p_1 > p_2$$

$$H_0^2: p_1 \geq p_2 \; vs \; H_a^2: p_1 < p_2$$

$$H_0^3: p_1 \leq p_3 \; vs \; H_a^3: p_1 > p_3$$

$$H_0^4: p_1 \geq p_3 \; vs \; H_a^4: p_1 < p_3$$

$$H_0^5: p_2 \leq p_3 \; vs \; H_a^5: p_2 > p_3$$

$$H_0^6: p_2 \geq p_3 \; vs \; H_a^6: p_2 < p_3$$

For the comparison of experimental treatment groups under Design C, the hypotheses equate to tests of H_0: $i_e \leq i_f \; vs \; H_a$: $i_e > i_f$ and H_0: $i_e \geq i_f \; vs \; H_a$: $i_e < i_f$ for all $e, f \in \{1,2, \ldots, K\}$ and $e \neq f$. As Shaffer's work is quite technical, condensing the results here would be insufficient. As a result, we instruct readers to consult the original publication, which contains full details of the testing procedure as well as examples of its implementation. It is then simple to understand its potential for application to the pairwise comparison of experimental treatments in a trial conducted under Design C methodology.

SECTION 6: FUTURE DIRECTIONS

The landscape of clinical trials has rapidly expanded, and trials evaluating the efficacy and/or safety of three or more treatments are now relatively common (Parmar, Carpenter, and Sydes, 2014). As such, a number of clinical trials aim to compare multiple experimental treatments to a single control. In this dissertation, we have provided an approach to conducting studies of this nature based upon waiting for a fixed number of events to occur in the control arm, leading to tests based on the negative multinomial distribution. This methodology represents an alternative approach to the multivariate extensions of Gail's Designs A and B (wait until a total number of events have occurred among the study arms or wait until a predetermined amount of time has elapsed, respectively), which are based on the multinomial distribution. We have provided methods for conducting one-sided global tests of treatment superiority and inferiority and combined these results to construct a two-sided test. Finally, we explored the possibility of comparing experimental treatments to each other using the work of Shaffer (1971). However, several open questions are implicated by the work in this dissertation and are discussed below.

Order Statistics of the Negative Multinomial Distribution

In Section 3, equations representing the distribution of discrete order statistics provided in Theorem 1 from Casella and Berger (2002), in conjunction with the comparative Poisson formulation of the negative multinomial distribution, were used to compute probabilities related to the order statistics of a balanced NMD. As one of the requirements of Theorem 1 is a sample of i.i.d. random variables, the theorem cannot be directly applied in the case of an unbalanced NMD, as the random variables are no longer

identically distributed. As a result, estimates of probabilities concerning order statistics of an unbalanced NMD were obtained via simulation. Though simulating a very large number of counts from an NMD provides sufficient accuracy for practical purposes, it remains of theoretical interest as to whether the comparative Poisson representation of the NMD can be used to provide a formula for the order statistics of an unbalanced NMD. In undertaking such work, it may be simplest to start by finding a representation for partially balanced NMDs, which constitute a subset of the unbalanced distributions.

One-sided Tests of Superiority and Inferiority

In Section 4, we presented the methodology for comparing multiple experimental treatments to a single control to determine if at least one experimental treatment had superior efficacy relative to the control. In Section 5, the corresponding test for treatment inferiority was presented. It is important to remember that the formulas for these tests of hypotheses were derived under the assumption of equal allocation of person years among the experimental and control treatment groups. However, to augment the utility of the tests, derivation of the formulas and updates to the *Des_Sup* and *Des_Inf* codes accommodating unequal allocation ratios is of interest.

Incorporating unequal allocation ratios is important for several reasons. Firstly, it is not always feasible to satisfy a $1:1:\ldots:1$ allocation ratio, even if such was intended, due to difficulties in recruiting certain patient populations. Secondly, depending on the specified alternative hypothesis, unequal allocation can reduce the total number of person years of follow-up or total number of subjects required for the trial (see Fleiss (1986), page 96). Finally, there may be a gain in statistical power in allocating a larger portion of study subjects to the control arm while keeping the portions allocated to each

experimental treatment arm equal. Equivalently, in some settings it has been shown that variance of the treatment effect estimate is minimized by allocating more person years of follow-up or subjects to the control arm, and such may be the case here (Dunnett, 1955 and Hoover and Blackwelder, 2001). This interest in unequal allocation ratios leads quite naturally to the subsequent objective of optimizing the allocation ratio to minimize the expected number of person years of follow-up until trial termination. Optimization of the allocation ratio will depend on whether the null or specified alternative hypothesis is assumed to be true.

The R functions *Des_Sup* and *Des_Inf* return the necessary parameters for designing a superiority or inferiority trial, respectively, under Design C. For a specified one-sided overall Type I error, a desired minimum level of pointwise power, and a given value of the rate ratio, these functions provide the number of control events which must be observed in the trial and the critical value for the test of hypothesis. Rather than determine the parameters needed to satisfy a required minimum power, researchers may be interested in determining the range of values of the rate ratio r under the alternative hypothesis which can be distinguished between the experimental treatment arms and the control arm at a given level of power. This question is of practical interest as the results of preclinical and early-phase studies are often used to inform researchers of appropriate parameter values for use in the design of large-scale trials.

Expected Person Years of Follow-up in Curtailed Design C Studies

Following the development of the methodology for conducting a test of superiority or inferiority, we considered the expected number of person years of follow-up until trial termination. As operating a trial under curtailed stopping rules can lead to a considerable

reduction in the total amount of person years and resources required for the study, algorithms were presented to estimate the expected number and standard deviation of person years until trial termination. The exact distribution for the number of person years was not provided as deriving this distribution may require enumeration of all possible ways the experimental treatment groups and the control group can acquire events leading to study stoppage. Though the simulated values are sufficiently accurate for practical use, obtaining the exact distribution for the number of person years in curtailed trials remains an open problem.

Stoppage at Interim Analysis Due to Futility

Another idea of interest is trial stoppage due to futility evaluated at one or more interim analyses. At an interim analysis, the conditional probability of rejecting the null hypothesis given the current collected data and assuming that the specified alternative hypothesis is true is computed (Snapinn et al., 2006). If this probability is high enough (based on a predetermined threshold), then the trial will continue. Otherwise, the trial is terminated early, again to "cut one's losses" and obtain a reduction in expenses. This differs from early termination due to curtailment, because under curtailment the trial is stopped early only once the ultimate decision is known (which is why the rejection regions under an uncurtailed and curtailed design coincide). Under futility stoppage, the decision made at the end of the trial is projected at an interim analysis using conditional probabilities, meaning that there is a non-zero probability that the decision to terminate the trial early at an interim analysis could be proven incorrect if the trial were to be continued to completion. As an example of futility under Design C, consider a clinical trial for superiority in which $K = 3$, $d_C = 30$, and $m = 18$. Suppose that one interim

analysis is conducted when the control group reaches 15 events, and at this time, the number of events observed in the three experimental treatment groups is $X_1 = 8, X_2 = 10$, and $X_3 = 7$. This example is illustrated below in Figure 8.

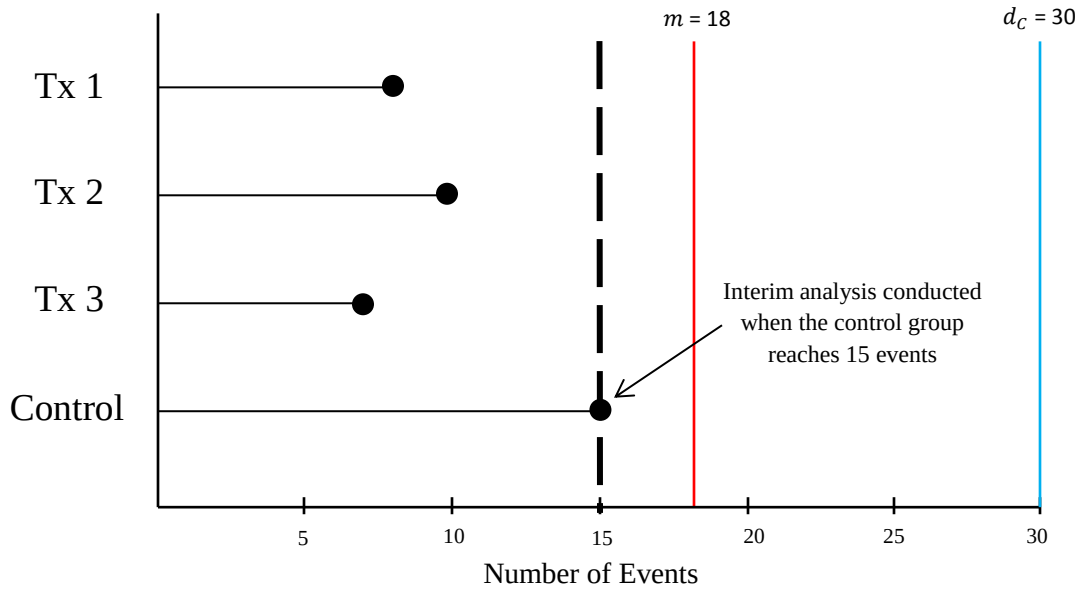


Figure 8: Example of a superiority trial under Design C with study parameters $K = 3$, $d_C = 30$, and $m = 18$ and one interim analysis conducted when the control group reaches 15 events

Suppose further that we will only proceed with the trial if there is greater than a 90% chance that the null hypothesis in Equation 4.2.1 will be rejected once the control group reaches 30 events based on the data collected at the interim analysis and assuming that the alternative hypothesis in Equation 4.2.1 is true; otherwise, we will terminate the trial at the interim analysis. To decide whether the trial will continue, we must calculate

$$P(at\ least\ one\ Tx\ group\ has \leq 18\ events\ when\ the\ control\ reaches\ 30\ events \mid X_1 = 8, X_2 = 10, X_3 = 7\ when\ the\ control\ reaches\ 15\ events)$$

If this probability is greater than 90%, then the trial will proceed until the control group reaches 30 events (i.e. the trial will proceed to completion). Otherwise, the trial will be

terminated early at the interim analysis as there is too great a chance that the trial will fail to reach significance at the end of the study. Computing the aforementioned conditional probabilities is a potential area of future research. Note that thus far, futility has been discussed in terms of terminating the entire trial if acceptance of the specified alternative hypothesis at the end of the study appears unlikely at interim analysis; we may also want to make decisions for individual treatment arms. For example, if at interim analysis in a superiority trial one or more treatment arms are projected to have too large a probability of reaching $m + 1$ events by the time the control group reaches d_C events , then follow-up in those individual arms would be terminated at the interim analysis. Returning to the example depicted in Figure 8, to make an individual decision for treatment 1 we would need to compute

$$P(X_1 \leq 18\ events\ when\ the\ control\ group\ reaches\ 30\ events \\ |X_1 = 8\ when\ the\ control\ group\ reaches\ 15\ events)$$

The event in the expression above is clearly characterized by a negative binomial process, and, using the marginal distribution of X_1, the probability may be calculated as $P(X_1 \leq 18 - 8 = 10)$ where $X_1 \sim NB\left(30 - 15 = 15, \frac{1}{1+r_1}\right)$. If this probability is sufficiently large, follow-up in treatment group 1 would continue until the control group reaches 30 events; otherwise, follow-up of treatment group 1 is terminated at the interim analysis as there is too great a chance that the treatment group will reach $m + 1 = 19$ events and the treatment found non-superior to the control at the time of stoppage at $d_C = 30$ events in the control group. Similar calculations would guide individual decisions for treatment groups 2 and 3.

As a final note on trial designs which incorporate futility analysis, futility can be combined with curtailment, and determining the futility bounds in this formulation would

also require additional work. An easily implemented closed-form solution for the expected number of person years of follow-up under a design which implements both curtailment and one or more interim analyses for assessment of futility may be difficult to achieve, but it may be possible to use simulation to estimate this value.

Power of Two-sided Hypothesis Tests

In Section 5, the one-sided tests of superiority and inferiority were combined to create a conservative two-sided hypothesis test. We did not discuss the overall power of the two-sided test due to difficulties in specifying appropriate definitions of power and deriving the corresponding power formulas; thus, power was selected individually for each direction (i.e. direction of treatment superiority and treatment inferiority). Defining and deriving the overall power of the two-sided test therefore remains an area of future work. A two-sided test based on the chi-squared distribution, first introduced by Stein (1955), was also described in Section 5. The power of this test was not found in the literature and was not addressed in this dissertation but may be of interest to other researchers.

Back-Up Approaches Under Failure to Reach d_C Events in a Trial Conducted Under Design C

Throughout this dissertation, it has been assumed that a clinical trial conducted under Design C will terminate once the control group reaches d_C events (unless curtailment is used in which case the trial may stop earlier once all experimental treatment groups have reached $m + 1$ or w events in a superiority or inferiority trial, respectively). However, in practice, investigators may need to stop a clinical trial prior to the control arm accumulating d_C events due to time or financial restrictions. This may occur if the true incidence rate of events for the control treatment is overestimated during study planning, resulting in events accumulating more slowly than anticipated during the trial; in such a

situation, we refer to the control group as being biased downwards. In this case, investigators will need to select a strategy to compare the experimental treatments to the control despite failure to reach d_C events in the control group. We propose two possible approaches:

1. Assume that conditional on the total number of events among all treatment groups (i.e. among all experimental treatment groups and the control group) at the time of stoppage, the distribution of $\tau, D_1, D_2, \ldots, D_K$ is multinomial, where $\tau < d_C$ is the number of events in the control group at the time of stoppage, and apply the multivariate version of Gail's Design A (see Section 2 and Hsu (2010) for a discussion of this test). Of course, the true distribution is not multinomial since the number of events in each experimental treatment group is dependent upon the rate at which the control group accumulates events (for example, if the control group accumulates events at a slower rate than anticipated, then under the global null hypothesis we would also expect the rate of event accrual in each experimental treatment group to be lower than anticipated).
2. Suppose that when the trial is stopped, $\tau < d_C$ events have occurred in the control arm, and let $\Delta = d_C - \tau$. Add Δ events to all experimental treatment arms and to the control arm, and conduct the test of hypothesis as if the control arm had actually reached d_C events.

We believe that both of the suggested ad-hoc approaches are conservative, meaning that the Type I error incurred will be no greater than the nominal error rate under which Design C was originally implemented, due to the discrete nature of the distributions involved. If a formal proof cannot be achieved, simulation may either refute this belief or

otherwise allow researchers to proceed under the assumption that the overall Type I error will be maintained when the above approaches are applied.

Appendix A: Proof that the Time to Obtain d_C Events in the Reference Group is Distributed $Gamma\left(d_C, \frac{1}{i_C}\right)$

The proof that the time to obtain d_C events in the reference group follows a Gamma distribution is well known[44] and is as follows: suppose that events accumulate in the reference group according to a Poisson process with parameter i_C. Let X denote the number of events that occur in the time interval $[0, t]$ and N_C the time until the $d_C{}^{\text{th}}$ event is observed. Then,

$$F_{N_C}(t) = P(N_C \leq t) = P(X \geq d_C) = 1 - P(X < d_C) = 1 - P(X \leq d_C - 1)$$

$$\Rightarrow F_{N_C}(t) = 1 - \sum_{x=0}^{d_C-1} e^{-i_C t} \frac{(i_C t)^x}{x!}$$

$$\Rightarrow f_{N_C}(t) = \frac{d}{dt} F_{N_C}(t) = -\left[\sum_{x=0}^{d_C-1} e^{-i_C t} \frac{x i_C (i_C t)^{x-1}}{x!} + \sum_{x=0}^{d_C-1} e^{-i_C t} \frac{-i_C (i_C t)^x}{x!}\right]$$

$$= i_C e^{-i_C t}\left[\sum_{x=0}^{d_C-1} \frac{(i_C t)^x}{x!} - \sum_{x=0}^{d_C-1} \frac{x (i_C t)^{x-1}}{x!}\right]$$

$$= i_C e^{-i_C t}\left[\sum_{x=0}^{d_C-1} \frac{(i_C t)^x}{x!} - \sum_{x=1}^{d_C-1} \frac{(i_C t)^{x-1}}{(x-1)!}\right] = i_C e^{-i_C t}\left[\sum_{x=0}^{d_C-1} \frac{(i_C t)^x}{x!} - \sum_{y=0}^{d_C-2} \frac{(i_C t)^y}{y!}\right]$$

$$= i_C e^{-i_C t}\left[\frac{(i_C t)^{d_C-1}}{(d_C-1)!}\right] = \frac{{i_C}^{d_C} t^{d_C-1} e^{-i_C t}}{\Gamma(d_C)} = \frac{1}{\Gamma(d_C)\left(\frac{1}{i_C}\right)^{d_C}} t^{d_C-1} e^{\frac{-t}{\left(\frac{1}{i_C}\right)}}$$

Hence, N_C is distributed $Gamma\left(d_C, \frac{1}{i_C}\right)$. This proof is also applied in Section 4.6 where it is used to show that the number of person years to reach $m + 1$ events in the k^{th} experimental treatment group is $Gamma\left(m + 1, \frac{1}{i_C}\right)$ or $Gamma\left(m + 1, \frac{1}{r_k i_C}\right)$

[44] See for example Casella, George, and Roger L. Berger. Statistical Inference. 1 ed., Duxbury Press, 1990.

corresponding to the null and alternative hypotheses in Equation 4.2.1, respectively. Assuming $i_C = 1$ yields $Gamma(m + 1,1)$ and $Gamma\left(m + 1, \frac{1}{r_k}\right)$ distributions, which were used in the simulation algorithms of Section 4.6.

Appendix B: Order Statistics of the Negative Multinomial Distribution when the Reference Outcome is Included

In Section 3, we considered the order statistics of the negative multinomial distribution. We derived formulas which provide exact probabilities when the distribution is balanced and provided R code to estimate the probabilities for unbalanced distributions. Here, we extend the definition of the order statistics of the NMD to include the reference outcome.

Again, consider rolling a die until we obtain five 6's (the reference outcome), and during the course of these trials we observe eight 1's, four 2's, five 3's, ten 4's, and seven 5's. Our definition of the order statistics in Section 3 indicated that $X_{(1)} = 4, X_{(2)} = 5, X_{(3)} = 7, X_{(4)} = 8$, and $X_{(5)} = 10$. Now, consider inclusion of the reference outcome (which is observed on a fixed number, $d_C = 5$, of trials) in the order statistics. In this case, we have $X_{(1)} = 4, X_{(2)} = 5, X_{(3)} = 5, X_{(4)} = 7, X_{(5)} = 8$, and $X_{(6)} = 10$.

It is clear from this example that when the reference outcome is included in the order statistics, it only disrupts the indexing of the original order statistics (the order statistics when the reference outcome is excluded) with value greater than d_C, increasing the index of each of these order statistics by one. This rule is predicated upon retaining the index of any original order statistic with value equal to d_C. For example, in the die experiment, the original second order statistic $X_{(2)}$ had value 5, which coincided with the value of d_C, and hence we retained the index "(2)" for this order statistic when considered in conjunction with the reference outcome, while the index of the original order statistics $X_{(3)}, X_{(4)}$, and $X_{(5)}$ (which each took on a value greater than d_C) each increased by one.

To calculate the probabilities of the order statistics when the reference outcome is included, we will use the following notation: let $p_{ji} = P(X_{(j)} = i)$ when the reference outcome is excluded and let $r_{ji} = P(X_{(j)} = i)$ when the reference outcome is included. As in Section 3, K is the number of comparator outcomes and d_C is the number of times the reference outcome is to be observed during the trials. From Equation 3.2.2., we know that for a balanced distribution

$$P(X_{(j)} \leq i) = \int_0^\infty \frac{x^{d_C-1}e^{-x}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left[\sum_{s=0}^{i} e^{-x}\frac{x^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i} e^{-x}\frac{x^s}{s!}\right]^{K-l} dx$$

which implies that

$$p_{ji} = P(X_{(j)} = i) = P(X_{(j)} \leq i) - P(X_{(j)} \leq i-1)$$

$$= \int_0^\infty \frac{x^{d_C-1}e^{-x}}{\Gamma(d_C)} \sum_{l=j}^{K} \binom{K}{l} \left\{ \left[\sum_{s=0}^{i} e^{-x}\frac{x^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i} e^{-x}\frac{x^s}{s!}\right]^{K-l} - \left[\sum_{s=0}^{i-1} e^{-x}\frac{x^s}{s!}\right]^l \left[1 - \sum_{s=0}^{i-1} e^{-x}\frac{x^s}{s!}\right]^{K-l} \right\} dx$$

When the distribution is unbalanced, the R function *unbalanced_order* can be used to estimate p_{ji}.

We now show how to compute r_{ji} from p_{ji}, making use of the fact that the number of trials resulting in the reference outcome is equal to the fixed value d_C.

For the minimum (i.e. $j = 1$), we have

$$r_{10} = p_{10}$$
$$r_{11} = p_{11}$$
$$\vdots$$
$$r_{1,d_C-1} = p_{1,d_C-1}$$
$$r_{1,d_C} = 1 - (p_{10} + p_{11} + \cdots + p_{1,d_C-1})$$
$$r_{1,d_C+1} = r_{1,d_C+2} = r_{1,d_C+3} = \cdots = 0$$

For the maximum (i.e. $j = K + 1$), we have

$$r_{K+1,0} = r_{K+1,1} = \cdots = r_{K+1,d_C-1} = 0$$
$$r_{K+1,d_C} = p_{K0} + p_{K1} + \cdots + p_{K,d_C-1} + p_{K,d_C} = 1 - \left(p_{K,d_c+1} + p_{K,d_c+2} + p_{K,d_c+3} + \cdots\right)$$
$$r_{K+1,d_C+1} = p_{K,d_c+1}$$
$$r_{K+1,d_C+2} = p_{K,d_c+2}$$
$$r_{K+1,d_C+3} = p_{K,d_c+3}$$
$$\vdots$$

Finally, for the remaining order statistics $j = 2, 3, \ldots, K$, we have

$$r_{j0} = p_{j0}$$
$$r_{j1} = p_{j1}$$
$$\vdots$$
$$r_{j,d_C-1} = p_{j,d_C-1}$$
$$r_{j,d_C} = 1 - \left(p_{j0} + p_{j1} + \cdots + p_{j,d_C-1} + p_{j-1,d_C+1} + p_{j-1,d_C+2} + p_{j-1,d_C+3} + \cdots\right)$$
$$r_{j,d_C+1} = p_{j-1,d_C+1}$$
$$r_{j,d_C+2} = p_{j-1,d_C+2}$$
$$r_{j,d_C+3} = p_{j-1,d_C+3}$$
$$\vdots$$

Appendix C: Conservativeness of the Test of Superiority when One or More Experimental Treatments are Inferior to the Control

In Equation 4.2.1, we assumed that $i_1 = i_2 = \cdots = i_K = i_C$ under the null hypothesis. Suppose that one or more of the experimental treatments are in fact inferior to the control. This can be represented by instead assuming that $i_1 = r_1 i_C, i_2 = r_2 i_C, \ldots, i_K = r_K i_C$ where $r_k \geq 1$ for $k = 1,2, \ldots, K$ and $r_k > 1$ for at least one of the r_k under the null hypothesis. We will show that in this case, the test of treatment superiority presented in Section 4.2 is conservative with respect to the overall Type I error; that is, we will show that the overall Type I error when one or more experimental treatments are inferior to the control does not exceed the overall Type I error when it is assumed that $i_1 = i_2 = \cdots = i_K = i_C$ under the null hypothesis. When one or more experimental treatments are inferior to the control, it is clear that the Type I error is given by

$$1 - \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right] \cdots \left[1 - \sum_{s=0}^{m} e^{-r_K t} \frac{(r_K t)^s}{s!}\right] dt$$

Hence, we must prove that

$$\begin{aligned} &1 - \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right] \cdots \left[1 - \sum_{s=0}^{m} e^{-r_K t} \frac{(r_K t)^s}{s!}\right] dt \\ &\qquad \leq 1 - \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-t} \frac{t^s}{s!}\right]^K dt \end{aligned} \tag{C1}$$

$$\begin{aligned} \Leftrightarrow &\int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-r_1 t} \frac{(r_1 t)^s}{s!}\right] \cdots \left[1 - \sum_{s=0}^{m} e^{-r_K t} \frac{(r_K t)^s}{s!}\right] dt \\ &\qquad \geq \int_0^\infty \frac{t^{d_C - 1} e^{-t}}{\Gamma(d_C)} \left[1 - \sum_{s=0}^{m} e^{-t} \frac{t^s}{s!}\right]^K dt \end{aligned}$$

From properties of the integral, we know that if $f(x) \geq g(x)$ for $a \leq x \leq b$, then $\int_a^b f(x)dx \geq \int_a^b g(x)\, dx$. Hence, we need to show

$$\left[1-\sum_{s=0}^{m} e^{-r_1 t}\frac{(r_1 t)^s}{s!}\right]\cdots\left[1-\sum_{s=0}^{m} e^{-r_K t}\frac{(r_K t)^s}{s!}\right] \geq \left[1-\sum_{s=0}^{m} e^{-t}\frac{t^s}{s!}\right]^K \quad \text{(C2)}$$

for $0 \leq t < \infty$. Suppose without loss of generality that $r_1 > 1$; we will show that

$$1-\sum_{s=0}^{m} e^{-r_1 t}\frac{(r_1 t)^s}{s!} \geq 1-\sum_{s=0}^{m} e^{-t}\frac{t^s}{s!}$$

$$\Leftrightarrow \sum_{s=0}^{m} e^{-r_1 t}\frac{(r_1 t)^s}{s!} \leq \sum_{s=0}^{m} e^{-t}\frac{t^s}{s!}$$

We know that $\sum_{s=0}^{m} e^{-x}\frac{x^s}{s!} = \int_x^\infty \frac{z^m e^{-z}}{\Gamma(m+1)}dz$ (see Casella and Berger (2002), page 130).

Therefore, we must show that

$$\int_{r_1 t}^{\infty} \frac{z^m e^{-z}}{\Gamma(m+1)}dz \leq \int_t^{\infty} \frac{z^m e^{-z}}{\Gamma(m+1)}dz \quad \text{(C3)}$$

Since $r_1 t \geq 0$ and $t \geq 0$ (since $0 \leq t < \infty$ for the integrals in (C1) and $r_k \geq 1$ for all k) and $m \geq 0$, we know that the integrands in (C3) are positive. Thus, since $r_1 t > t$ (since $r_1 > 1$ by assumption), we know that

$$\int_{r_1 t}^{\infty} \frac{z^m e^{-z}}{\Gamma(m+1)}dz < \int_t^{\infty} \frac{z^m e^{-z}}{\Gamma(m+1)}dz$$

This obviously implies that (C3) holds. It should be clear that if any number of the experimental treatments are inferior to the control (i.e. if any subset of the r_k are strictly greater than one), then the inequality in (C2) will hold. This completes the proof of the relationship specified in (C1).

Similar computations will show that the inferiority test presented in Section 5.1 is conservative with respect to overall Type I error when one or more experimental treatments are superior to the control under the null hypothesis. Finally, as the two-sided test presented in Section 5.2 is based on the rejection boundaries obtained from the one-

sided tests of superiority and inferiority, the two-sided test is conservative since the one-sided tests are conservative.

Appendix D: Alternate Formulas for the Minimum and Maximum Number of Events in a Negative Multinomial Distribution

In Section 3, we considered the order statistics of the negative multinomial distribution, providing an explicit formula in the case of balanced distributions, and in Section 4 we derived simple expressions for the minimum and maximum. Here we draw attention to some relevant formulas derived by Olkin and Sobel (1965) and Joshi (1972). Olkin and Sobel consider a negative multinomial design in which there are $k+2$ mutually exclusive cells denoted by $C_1, \dots, C_{k+2}$ and observations are recorded until cell C_{k+1} contains s observations. They consider the events E_1 and E_2 described below:

$E_1: at\ the\ time\ of\ stopping, X_j \geq s_j\ for\ all\ j = 1, \dots, k$
$E_2: at\ the\ time\ of\ stopping, X_j \leq s_j - 1\ for\ all\ j = 1, \dots, k$

where the s_j are non-negative integers and X_j denotes the number of observations in cell C_j at the time of stopping for $j = 1, \dots, k$. Letting $x_0 = \sum_{i=1}^{k} x_i$, the corresponding formula for $P\{E_1\}$ is

$$P\{E_1\} = \sum_{\alpha=s}^{\infty} \sum_{x_1=s_1}^{\infty} \cdots \sum_{x_k=s_k}^{\infty} \frac{\Gamma(\alpha)\left(\prod_{i=1}^{k} p_i^{x_i}\right) p_{k+1}^{s} p_{k+2}^{\alpha-s-x_0}}{\Gamma(s)\Gamma(\alpha - s - x_0 + 1)\left(\prod_{i=1}^{k} x_i!\right)}$$

where p_i is the probability of observing cell $C_i, i = 1, \dots, k+2$. Clearly, if the s_j are all equal to a common value, then E_1 corresponds to the minimum and E_2 corresponds to the maximum of $(X_1, \dots, X_k)$. The authors provide additional equivalent formulas for $P\{E_1\}$ throughout the paper, but the formulas are unwieldy, particularly for the purposes of this dissertation. Joshi improves upon Olkin and Sobel's results by utilizing the comparative Poisson representation of the NMD to determine $P_X(\boldsymbol{x}) = P(X_i \leq x_i, i = 1,2, \dots, s)$ as follows:

$$P_X(\boldsymbol{x}) = \int_0^{\infty} \prod_{i=1}^{s} \sum_{r_i=0}^{x_i} \frac{e^{-\lambda_i\theta}(\lambda_i\theta)^{r_i}\theta^{k-1}e^{-\theta}}{r_i!\,\Gamma(k)}\,d\theta$$

$$= \int_0^{\infty} \prod_{i=1}^{s} \int_{\lambda_i\theta}^{\infty} \frac{e^{-z_i}z_i^{x_i}}{x_i!}\,dz_i \frac{e^{-\theta}\theta^{k-1}}{\Gamma(k)}\,d\theta = P\left(\frac{Z_i}{\Theta} > \lambda_i, i = 1,2,\dots,s\right)$$

where $Z_1, Z_2, \dots, Z_s, \Theta$ are mutually independent gamma random variables with density functions $f_{Z_i}(z) = \frac{z^{x_i}e^{-z}}{x_i!}$ and $f_{\Theta}(\theta) = \frac{\theta^{k-1}e^{-\theta}}{\Gamma(k)}$. Similar methods can be employed to find $Q_X(\boldsymbol{x}) = P(X_i > x_i, i = 1,2,\dots,s)$. When the x_i are all equal (say $x_i = a$ for all i), Joshi's results can be applied to determine $P(\min(\mathrm{X}_1, \dots, X_s) \leq a)$ and $P(\max(\mathrm{X}_1, \dots, X_s) \leq a)$, though Joshi does not provide the explicit results for the minimum and maximum as we have in Equations 4.2.4 and 4.2.5, respectively.

Appendix E: R Functions Used in this Dissertation[45]

1. *balanced_order_less(dc,j,i,K)* and ***balanced_order_equal(dc,j,i,K)*:**
balanced_order_less takes the arguments d_C (number of trials resulting in the reference outcome to be observed), j (denotes the j^{th} order statistic), i (takes a value in 0,1,2,…), and K (number of comparator outcomes in the experiment, i.e. not including the reference outcome). For a balanced negative multinomial distribution, *balanced_order_less* returns $P(X_{(j)} \leq i)$ when there are K comparator outcomes and trials are conducted until d_C reference outcomes are observed. The function *balanced_order_equal* takes the same arguments as *balanced_order_less* and returns $P(X_{(j)} = i)$.

***balanced_order_less* Code:**

```
balanced_order_less<-function(dc,j,i,K){
if(dc<=0|dc%%1!=0|j<=0|j%%1!=0|j>K|i<0|i%%1!=0|K<=0|K%%1!=0){
cat("Error. Valid values of arguments are as follows:\n")
cat("dc must be a positive integer\n")
cat("j must be an integer from 1 to",K,"inclusive\n")
cat("i must be an integer value greater than or equal to 0\n")
cat("K must be a positive integer\n")
}
else{
empty<-c()
for(l in j:K){
terms<-function(x){dgamma(x,dc)*choose(K,l)*((ppois(i,x))^l)*((1-ppois(i,x))^(K-l))}
int_val<-integrate(terms,0,Inf)$value
empty<-c(empty,int_val)
}
prob_val<-sum(empty)
return(prob_val)
}
}
```

***balanced_order_equal* Code:**

```
balanced_order_equal<-function(dc,j,i,K){
if(dc<=0|dc%%1!=0|j<=0|j%%1!=0|j>K|i<0|i%%1!=0|K<=0|K%%1!=0){
cat("Error. Valid values of arguments are as follows:\n")
cat("dc must be a positive integer\n")
cat("j must be an integer from 1 to",K,"inclusive\n")
cat("i must be an integer value greater than or equal to 0\n")
cat("K must be a positive integer\n")
}
```

[45] All results in this dissertation from functions which depend upon simulation were generated using seed value 1234567 in R version 3.2.0 and using 100,000 simulations, unless otherwise indicated.

```
else{
p_equal<-balanced_order_less(dc,j,i,K)-balanced_order_less(dc,j,i-1,K)
return(p_equal)
}
}
```

Examples:

(1) For a balanced negative multinomial distribution in which there are 5 comparator outcomes and the number of trials resulting in the reference outcome to be observed is 10, the probability that the fourth order statistic is less than or equal to 4 is given by:

```
balanced_order_less(10,4,4,5)
0.01403157
```

(2) Under the settings in example (1), the probability that the fourth order statistic is equal to 4 is given by:

```
balanced_order_equal(10,4,4,5)
0.01031401
```

2. *unbalanced_order(probs,dc,j,i,sims)*: *unbalanced_order* takes the arguments *probs* (vector of length K, where K is the number of comparator outcomes in the NMD, containing the probabilities of a trial resulting in each comparator outcome, i.e. not including the reference outcome), d_C (number of trials resulting in the reference outcome to be observed), j (denotes the j^{th} order statistic), i (takes a value in 0,1,2,…), and *sims* (number of simulations used to estimate the probability). *unbalanced_order* returns an estimate of $P(X_{(j)} \leq i)$ for unbalanced negative multinomial distributions. This function can also be used to find the probability that the j^{th} order statistic is equal to i by computing $P\left(X_{(j)} \leq i\right) - P(X_{(j)} \leq i-1)$ (see example (2) below for an illustration).

***unbalanced_order* Code:**

```
library(MGLM)
unbalanced_order<-function(probs,dc,j,i,sims){
if(any(probs<=0)){
cat("Error: All entries in probs vector must be greater than 0\n")
}
else if(sum(probs)>=1){
cat("Error: Sum of probabilities in probs vector may not be greater than or equal to 1\n")
cat("probs vector should contain the probabilities of an outcome in each comparator
group\n")
cat("(i.e. not including the probability of an outcome in the reference group)\n")
}
```

```
else
if(dc<=0|dc%%1!=0|j<=0|j%%1!=0|j>length(probs)|i<0|i%%1!=0|sims<=0|sims%%1!=0)
{
cat("Error. Valid values of arguments are as follows:\n")
cat("dc must be a positive integer\n")
cat("j must be an integer from 1 to",length(probs),"inclusive, based on the vector of
probabilities entered\n")
cat("i must be an integer value greater than or equal to 0\n")
cat("sims must be a positive integer\n")
}
else{
vec<-rnegmn(sims,probs,dc)
new<-t(apply(vec,1,sort))
column<-new[,j]
emp<-c()
for(val in column){
if(val<=i){
emp<-c(emp,1)
}
}
final<-sum(emp)/sims
return(final)
}
}
```

Examples:

(1) For an unbalanced negative multinomial distribution in which there are 5 comparator outcomes with probabilities 0.1, 0.1, 0.3, 0.2, and 0.1 (so the probability the reference outcome is observed in a trial is 0.2) and the number of reference outcomes to be observed is 10, the probability that the third order statistic is less than or equal to 4 using 1,000,000 simulations and a seed value of 1234567 is estimated to be:

```
unbalanced_order(c(.1,.1,.3,.2,.1),10,3,4,1000000)
0.218617
```

(2) Under the settings in example (1), the estimated probability that the third order statistic is equal to 4 using 1,000,000 simulations and a seed value of 1234567 is estimated to be:

```
unbalanced_order(c(.1,.1,.3,.2,.1),10,3,4,1000000)-
unbalanced_order(c(.1,.1,.3,.2,.1),10,3,3,1000000)
0.119906
```

3. *Des_Sup(K,alpha,r,pwr)*: *Des_Sup* takes the arguments *K* (number of experimental treatment groups, i.e. not including the control group), *alpha* (nominal one-sided overall

Type I error at which the test of hypothesis is to be conducted), r (estimate of the rate ratio of the experimental treatment group to the control group which we wish to detect), and pwr (minimum desired pointwise power of the study). *Des_Sup* returns the number of events d_C to be observed in the control group, the critical value m for the hypothesis test, the true overall Type I error achieved, and the true pointwise power achieved in a superiority trial conducted under Design C methodology. *Des_Sup* can also be used to generate the corresponding results under the Bonferroni method by setting the number of experimental treatment groups equal to one and replacing the nominal overall Type I error with α_{ovr}/K (see example (2) below for an illustration).

Des_Sup calls several functions to compute the results. Among these is the function *Prob*, which takes the arguments d_C (number of events to observe in the control group), m (an integral value), and *K* (number of experimental treatment groups) and returns $P(\min(D_1, \ldots, D_K) \leq m)$ under the null hypothesis in Equation 4.2.1. This function can be used to find the p-value for a test of treatment superiority by substituting the observed minimum number of events among the experimental treatment groups for m (see example (3) below for an illustration).

***Des_Sup* Code:**

```
Prob<-function(dc,m,K){
new_function<-function(x){dgamma(x,dc)*((1-ppois(m,x))^K)}
result<-1-integrate(new_function,0,Inf)$value
return(result)
}

Control_ind<-function(K,alpha){
counter<-1
while(Prob(counter,0,K)>alpha){
counter<-counter+1
}
control_start<-counter
return(control_start)
}

CritVal<-function(dc,K,alpha){
ind<-0
while((Prob(dc,ind,K)<alpha)&(Prob(dc,ind+1,K)<=alpha)){
ind<-ind+1
}
x<-ind
return(x)
}

PointPwr<-function(dc,K,alpha,r){
s_comp<-c()
```

```
for(term in 0:(CritVal(dc,K,alpha))){
component<-(choose(dc+term-1,term)*(r^term))/((1+r)^(dc+term))
s_comp<-c(s_comp,component)
}
power<-sum(s_comp)
return(power)
}

Des_Sup<-function(K,alpha,r,pwr){
if(K<=0|K%%1!=0|alpha<=0|alpha>=1|r<=0|r>1|pwr<=0|pwr>=1){
cat("Error. Valid values of arguments are as follows:\n")
cat("K must be a positive integer\n")
cat("0<alpha<1\n")
cat("0<r<=1\n")
cat("0<pwr<1\n")
}
else{
start<-Control_ind(K,alpha)
while(PointPwr(start,K,alpha,r)<pwr){
start<-start+1
}
cat("The number of control group events dc is",start)
cat("\nThe critical value m is",CritVal(start,K,alpha))
cat("\nThe true overall Type I error is",Prob(start,CritVal(start,K,alpha),K))
cat("\nThe true pointwise power is",PointPwr(start,K,alpha,r),"\n")
}
}
```

Examples:

(1) The design parameters for a superiority trial in which four experimental treatment groups are compared to a control group at nominal one-sided overall Type I error of 0.05, desired pointwise power equal to 0.9, and a rate ratio r of 0.2 are:

```
Des_Sup(4,.05,.2,.9)
The number of control group events dc is 18
The critical value m is 6
The true overall Type I error is 0.03944082
The true pointwise power is 0.9088288
```

(2) The Bonferroni design parameters for the superiority trial described in example (1) are:

```
Des_Sup(1,.05/4,.2,.9)
The number of control group events dc is 18
The critical value m is 6
```

The true overall Type I error is 0.01132792
The true pointwise power is 0.9088288

(3) The p-value for the superiority test corresponding to the trial described in example (1) when the minimum number of events observed among the experimental treatment groups is 3 is:

Prob(18,3,4)
0.002885246

4. *Des_Inf(K,alpha,r,pwr)*: *Des_Inf* takes the arguments K (number of experimental treatment groups, i.e. not including the control group), *alpha* (nominal one-sided overall Type I error at which the test of hypothesis is to be conducted), r (estimate of the rate ratio of the experimental treatment group to the control group which we wish to detect), and pwr (minimum desired pointwise power of the study). *Des_Inf* returns the number of events d_C to be observed in the control group, the critical value w for the hypothesis test, the true overall Type I error achieved, and the true pointwise power achieved in an inferiority trial conducted under Design C methodology. Values under the Bonferroni approach and the p-value for the test of hypothesis can be found in the same manner as was explained for the *Des_Sup* routine above.

***Des_Inf* Code:**

```
Prob<-function(dc,w,K){
new_function<-function(x){dgamma(x,dc)*((ppois(w-1,x))^K)}
result<-1-integrate(new_function,0,Inf)$value
return(result)
}

PointPwr<-function(dc,w,r){
s_comp<-c()
for(term in 0:w-1){
component<-(choose(dc+term-1,term)*(r^term))/((1+r)^(dc+term))
s_comp<-c(s_comp,component)
}
power<-1-sum(s_comp)
return(power)
}

Des_Inf<-function(K,alpha,r,pwr){
if(K<=0|K%%1!=0|alpha<=0|alpha>=1|r<1|pwr<=0|pwr>=1){
cat("Error. Valid values of arguments are as follows:\n")
cat("K must be a positive integer\n")
cat("0<alpha<1\n")
cat("r>=1\n")
cat("0<pwr<1\n")
```

```
}
else{
start<-1
ind<-0
while(start>0){
while(Prob(start,ind,K)>alpha){
ind<-ind+1
}
if(PointPwr(start,ind,r)>=pwr){
cat("The number of control group events dc is",start)
cat("\nThe critical value w is",ind)
cat("\nThe true overall Type I error is",Prob(start,ind,K))
cat("\nThe true pointwise power is",PointPwr(start,ind,r))
cat("\n")
start<-0
}
else{
start<-start+1
}
}
}
}
```

Examples:

(1) The design parameters for an inferiority trial in which four experimental treatment groups are compared to a control group at nominal one-sided overall Type I error equal to 0.05, desired pointwise power of 0.8, and rate ratio r equal to 2 are:

```
Des_Inf(4,.05,2,.8)
The number of control group events dc is 30
The critical value w is 49
The true overall Type I error is 0.04866245
The true pointwise power is 0.8008007
```

5. *Null_Time(K,dc,crit,test,sims)*: *Null_Time* takes the arguments K (number of experimental treatment groups, i.e. not including the control group), dc (number of events to be observed in the control group), *crit* (the critical value for the test of hypothesis, i.e. m for a superiority trial or w for an inferiority trial), *test* (either "Sup" or "Inf" to specify whether the trial is of treatment superiority or inferiority, respectively), and *sims* (number of simulations used to estimate the person year values). *Null_Time* returns the estimated expected number, standard deviation, and 95% empirical confidence interval of person years of follow-up until trial termination (assuming $i_C = 1$) for either a fully curtailed superiority trial or a fully curtailed inferiority trial under the null hypothesis in Equation 4.2.1 or Equation 5.1.1, respectively.

***Null_Time* Code:**

```
ExpTime<-function(K,dc,crit,test){
if(toString(test)=="Sup"){
Txs<-rgamma(K,crit+1,1)
}
else if(toString(test)=="Inf"){
Txs<-rgamma(K,crit,1)
}
Cont<-rgamma(1,dc,1)
Txstar<-Txs
Txstar[Txstar>=Cont]<-Cont
ntkstar<-Txstar
Contstar<-Cont
ifelse(Cont>max(Txs),Contstar<-max(Txs),Contstar<-Cont)
sumall<-sum(ntkstar,Contstar)
return(sumall)
}

Null_Time<-function(K,dc,crit,test,sims){
v<-c("Sup","Inf")
if(dc<=0|dc%%1!=0|crit<0|crit%%1!=0|sims<0|sims%%1!=0|K<=0|K%%1!=0){
cat("Error. Valid values of arguments are as follows:\n")
cat("K must be a positive integer\n")
cat("dc must be a positive integer\n")
cat("crit must be an integer value greater than or equal to 0\n")
cat("sims must be a positive integer\n")
}
else if(is.element(toString(test),v)==FALSE){
cat("Error: Must specify 'Sup' or 'Inf' as an argument\n")
}
else{
Times<-vector()
for(i in 1:sims){
newval<-ExpTime(K,dc,crit,test)
Times<-c(Times,newval)
}
estimateTime<-mean(Times)
estimatestd<-sd(Times)
Sorted<-sort(Times)
TwoFive<-Sorted[.025*sims]
NineSeven<-Sorted[.975*sims]
cat("The estimated time is",estimateTime,"\n")
cat("The estimated standard deviation is",estimatestd,"\n")
cat("Interval based on 2.5 and 97.5 percentiles is (",TwoFive,",",NineSeven,")\n",sep="")
}
```

}

Examples:

(1) The estimated mean, standard deviation, and 95% empirical confidence interval of person years of follow-up until trial termination for a fully curtailed superiority trial under the null hypothesis in Equation 4.2.1 when there are four experimental treatment groups, the number of events to observe in the control group is $d_C = 25$, and the critical value is $m = 8$ using 100,000 simulations and seed value 1234567 are:

Null_Time(4,25,8,"Sup",100000)
The estimated time is 48.21195
The estimated standard deviation is 8.148895
Interval based on 2.5 and 97.5 percentiles is (33.51516,65.38062)

(1) The estimated mean, standard deviation, and 95% empirical confidence interval of person years of follow-up until trial termination for a fully curtailed inferiority trial under the null hypothesis in Equation 5.1.1 when there are four experimental treatment groups, the number of events to observe in the control group is $d_C = 30$, and the critical value is $w = 49$ using 100,000 simulations and seed value 1234567 are:

Null_Time(4,30,49,"Inf",100000)
The estimated time is 149.8836
The estimated standard deviation is 27.05042
Interval based on 2.5 and 97.5 percentiles is (100.9101,206.8324)

6. *Alt_Time(dc,crit,vec,test,sims)*: *Alt_Time* takes the arguments dc (number of events to be observed in the control group), *crit* (the critical value for the test of hypothesis, i.e. m for a superiority trial or w for an inferiority trial), *vec* (a vector of length equal to the number of experimental treatment groups with entries corresponding to the rate ratios of each experimental treatment group to the control group, i.e. a vector of the form $c(r_1, r_2, \dots, r_K)$), *test* (either "Sup" or "Inf" to specify whether the trial is of treatment superiority or inferiority, respectively), and *sims* (number of simulations used to estimate the person year values). *Alt_Time* returns the estimated expected number, standard deviation, and 95% empirical confidence interval of person years of follow-up until trial termination (assuming $i_C = 1$) for either a fully curtailed superiority trial or a fully curtailed inferiority trial under the alternative hypothesis in Equation 4.2.1 or 5.1.1, respectively.

***Alt_Time* Code:**

```
ExpTime_Alt<-function(dc,crit,vec,test){
empty<-c()
if(toString(test)=="Sup"){
for(i in vec){
randv<-rgamma(1,crit+1,i)
```

```
empty<-c(empty,randv)
}
}
else if(toString(test)=="Inf"){
for(i in vec){
randv<-rgamma(1,crit,i)
empty<-c(empty,randv)
}
}
Cont<-rgamma(1,dc,1)
Txstar<-empty
Txstar[Txstar>=Cont]<-Cont
ntkstar<-Txstar
Contstar<-Cont
ifelse(Cont>max(empty),Contstar<-max(empty),Contstar<-Cont)
sumall<-sum(ntkstar,Contstar)
return(sumall)
}

Alt_Time<-function(dc,crit,vec,test,sims){
v<-c("Sup","Inf")
if(dc<=0|dc%%1!=0|crit<0|crit%%1!=0|sims<0|sims%%1!=0){
cat("Error. Valid values of arguments are as follows:\n")
cat("dc must be a positive integer\n")
cat("crit must be an integer value greater than or equal to 0\n")
cat("sims must be a positive integer\n")
}
else if(any(vec<=0)){
cat("Error. All entries in vec must be greater than 0\n")
}
else if(is.element(toString(test),v)==FALSE){
cat("Error: Must specify 'Sup' or 'Inf' as an argument\n")
}
else{
Alt_Times<-vector()
for(i in 1:sims){
newval<-ExpTime_Alt(dc,crit,vec,test)
Alt_Times<-c(Alt_Times,newval)
}
estimateTime<-mean(Alt_Times)
estimatestd<-sd(Alt_Times)
Sorted<-sort(Alt_Times)
TwoFive<-Sorted[.025*sims]
NineSeven<-Sorted[.975*sims]
cat("The estimated time is",estimateTime,"\n")
cat("The estimated standard deviation is",estimatestd,"\n")
```

```
cat("Interval based on 2.5 and 97.5 percentiles is (",TwoFive,",",NineSeven,")\n",sep="")
}
}
```

Examples:

(1) The estimated mean, standard deviation, and 95% empirical confidence interval of person years of follow-up until trial termination for a fully curtailed superiority trial under the alternative hypothesis in Equation 4.2.1 when there are four experimental treatment groups, the number of events to observe in the control group is $d_C = 25$, the critical value is $m = 8$, and the anticipated rate ratios in the four experimental treatment groups are 0.2, 0.5, 0.6, and 0.4 using 100,000 simulations and seed value 1234567 are:

```
Alt_Time(25,8,c(.2,.5,.6,.4),"Sup",100000)
The estimated time is 101.201
The estimated standard deviation is 14.86817
Interval based on 2.5 and 97.5 percentiles is (73.64713,131.8204)
```

(2) The estimated mean, standard deviation, and 95% empirical confidence interval of person years of follow-up until trial termination for a fully curtailed inferiority trial under the alternative hypothesis in Equation 5.1.1 when there are four experimental treatment groups, the number of events to observe in the control group is $d_C = 30$, the critical value is $w = 49$, and the anticipated rate ratios in the four experimental treatment groups are 5, 2, 10, and 2 using 100,000 simulations and seed value 1234567 are:

```
Alt_Time(30,49,c(5,2,10,2),"Inf",100000)
The estimated time is 87.72808
The estimated standard deviation is 7.712449
Interval based on 2.5 and 97.5 percentiles is (72.58614,102.9377)
```

References

Bates, Grace, and Jerzy Neyman. "Contributions to the Theory of Accident Proneness I. An Optimistic Model of the Correlation Between Light and Severe Accidents." University of California publications in statistics, vol. 1, no. 9, 30 Apr. 1952, pp. 215-53.

Birnbaum, Allan. "Some Procedures for Comparing Poisson Processes or Populations." Biometrika, vol. 40, no. 3/4, Dec. 1953, pp. 447-49.

Birnbaum, Allan. "Statistical Methods for Poisson Processes and Exponential Populations." Journal of the American Statistical Association, vol. 49, no. 266, June 1954, pp. 254-66.

Brown, Charles C., and Sylvan B. Green. "Additional Power Computations for Designing Comparative Poisson Trials." American Journal of Epidemiology, vol. 115, no. 5, 1982, pp. 752-58.

Casella, George, and Roger L. Berger. Statistical Inference. 2nd ed., Duxbury Press, 2002, pp. 66-230.

Chatelain, Florent, Sophie Lambert-Lacroix, and Jean-Yves Tourneret. "Pairwise likelihood estimation for multivariate mixed Poisson models generated by Gamma intensities." Statistics and Computing, vol. 19, no. 3, 2009, pp. 283-301.

Dunnett, Charles W. "A Multiple Comparison Procedure for Comparing Several Treatments with a Control." Journal of the American Statistical Association, vol. 50, no. 272, Dec. 1995, pp. 1096-121.

Engel, J. "Split-Plot Design: Model and Analysis for Count Data." Statistica Neerlandica, vol. 40, no. 1, 1986, pp. 21-33.

Ferrari, A, G Letac, and J Y. Tourneret. "Multivariate Mixed Poisson Distributions." Signal Processing Conference, Sep. 2004, pp. 1067-70.

Fleiss, Joseph L. The design and analysis of clinical experiments. Wiley, 1986, p. 96.

Frey, Sharon, Timo Vesikari, Agnieszka Szymczakiewicz-Multanowska, Maria Lattanzi, Allen Izu, Nicola Groth, and Sandra Holmes. "Clinical Efficacy of Cell Culture-Derived and Egg-Derived Inactivated Subunit Influenza Vaccines in Healthy Adults." Clinical Infectious Diseases, vol. 51, no. 9, 1 Nov. 2010, pp. 997-1004.

Gail, Mitchell. "Power Computations for Designing Comparative Poisson Trials." Biometrics, vol. 30, no. 2, June 1974, pp. 231-37.

Gallager, Robert G. Discrete Stochastic Processes. Boston, Kluwer Academic Publishers, 1996, pp. 31-55.

Greenwood, Major, and G. Udny Yule. "An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents." Journal of the Royal Statistical Society, vol. 83, no. 2, Mar. 1920, pp. 255-79.

Hoover, Donald R., and William C. Blackwelder. "Allocation of subjects to test null relative risks smaller than one." Statistics in Medicine, vol. 20, 30 Oct. 2001, pp. 3071-82.

Hsu, Tzu-Lin. Comparative Poisson trials for comparing multiple new treatments to the control. PhD Dissertation. Rutgers University, 2010. Retrieved from https://doi.org/doi:10.7282/T3FJ2GHX

Johnson, Norman L., Samuel Kotz, and Adrienne W. Kemp. *Univariate Discrete Distributions*. 2nd ed., New York, John Wiley & Sons, Inc., 1992, pp. 199-235.

Johnson, Norman L., Samuel Kotz, and N Balakrishnan. *Discrete Multivariate Distributions*. New York, John Wiley & Sons, Inc., 1997, pp. 31-123.

Joshi, S W. "Integral Expressions for Tail Probabilities of the Negative Multinomial Distribution." Annals of the Institute of Statistical Mathematics, vol. 27, no. 1, 1975, pp. 95-97.

Karlis, Dimitris, and Evdokia Xekalaki. "Mixed Poisson Distributions." International Statistical Review, vol. 73, no. 1, 2005, pp. 35-58.

Le Gall, Françoise. "The modes of a negative multinomial distribution." Statistics and Probability Letters, vol. 76, 2006, pp. 619-24.

Lehmann, E.L., and H Scheffé. "Completeness, similar regions, and unbiased estimation: Part II." Sankhyā, vol. 15, 1955, pp. 219-36.

Lehmann, E.L., and Joseph P. Romano. Testing Statistical Hypotheses. 3rd ed., New York, Springer, 2005, pp. 124-350.

Neyman, Jerzy. "Certain Chance Mechanisms Involving Discrete Distributions." Sankhyā: The Indian Journal of Statistics, Series A, vol. 27, no. 2/4, Dec. 1965, pp. 249-58.

Nguyen, Truc T., Arjun K. Gupta, Diem M. Nguyen, and Yining Wang. "Characterizations of Negative Multinomial Distributions Based on Conditional Distributions." Metrika, vol. 66, no. 3, Nov. 2007, pp. 315-22.

Olkin, Ingram, and Milton Sobel. "Integral Expressions for Tail Probabilities of the Multinomial and Negative Multinomial Distributions." Biometrika, vol. 52, no. 1/2, June 1965, pp. 167-79.

Ord, J.K., G.P. Patil, and C. Taillie, editors. Statistical distributions in ecological work. Vol. 4, Burtonsville, International Co-operative Publishing House, 1979, pp. 157-91.

Papageorgiou, H. "A Characterization of the Negative Multinomial as a Poisson Mixture." Journal of Statistical Computation and Simulation, vol. 17, 1983, pp. 245-47.

Parmar, Mahesh K.B., James Carpenter, and Matthew R. Sydes. "More multiarm randomised trials of superiority are needed." Lancet, vol. 384, no. 9940, 26 July 2014, pp. 283-84.

Patil, G P., and M T. Boswell. "On a Class of Multivariate Birth and Death Processes." Sankhyā: The Indian Journal of Statistics, vol. 34, no. 3, Sept. 1972, pp. 293-96.

Peng, Jie, and K Krishnamoorthy. "Conditional and Unconditional Tests for Comparing Several Poisson Means." Journal of Applied Statistical Science, vol. 18, no. 3, 2010, pp. 319-26.

Przyborowski, J, and H Wilenski. "Homogeneity of results in testing samples from Poisson series." Biometrika, vol. 31, no. 3-4, 1 March 1940, pp. 313-23.

Shaffer, Juliet. "An Exact Multiple Comparisons Test for a Multinomial Distribution." British Journal of Mathematical and Statistical Psychology, vol. 24, no. 2, 1971, pp. 267-72.

Sibuya, Masaaki, Isao Yoshimura, and Ryoichi Shimizu. "Negative Multinomial Distribution." Annals of the Institute of Statistical Mathematics, vol. 16, no. 3, 1964, pp. 409-26.

Singh, Bahadur. "Bayesian Inference for Comparing Several Treatment Means Under a Poisson Model." Communications in Statistics, A: Theory and Methods, vol. 9, 1980, pp. 1137-45.

Sinoquet, Hervé, and Raymond Bonhomme. "A Theoretical Analysis of Radiation Interception in a Two-Species Plant Canopy." Mathematical Biosciences, vol. 105, 1991, pp. 23-45.

Snapinn, Steven, Mon-Gy Chen, Qi Jiang, and Tony Koutsoukos. "Assessment of futility in clinical trials." Pharmaceutical Statistics, vol. 5, no. 4, 2006, pp. 273-81.

Steyn, H S., N.A.S. Crowther, E L. Raath, and T M. Shaw. "A General Model for Negative Multinomial Frequency Counts." South African Statistical Journal, vol. 23, no. 1, Jan. 1989, pp. 85-117.

Steyn, H S. "On Discrete Multivariate Probability Functions of Hypergeometric Type." Indagationes Mathematicae (Proceedings), vol. 58, 1955, pp. 588-95.

Steyn, H S. "On χ^2-tests for contingency tables of negative multinomial types." Statistica Neerlandica, vol. 13, no. 4, 1959, pp. 433-44.

Suissa, Samy, and Rachid Salmi. "Unidirectional Multiple Comparisons of Poisson Rates." Statistics in Medicine, vol. 8, 1989, pp. 757-64.

Tsui, Kam-Wah. "Multiparameter Estimation for Some Multivariate Discrete Distributions with Possibly Dependent Components." Annals of the Institute of Statistical Mathematics, vol. 38, no. 1, 1986, pp. 45-56.

Zhang, Yiwen, and Hua Zhou (2017). MGLM: Multivariate Response Generalized Linear Models. R package version 0.0.8. URL https://CRAN.R-project.org/package=MGLM

Zhou, Hua, and Kenneth Lange. "MM Algorithms for Some Discrete Multivariate Distributions." Journal of Computational and Graphical Statistics, vol. 19, no. 3, 1 Sept. 2010, pp. 645-65.