Tutorial in Biostatistics

Statistics
in Medicine

Received 14 December 2010, Accepted 1 March 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5381

Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data

Steven A. Julious*† and Michael J. Campbell

This article gives an overview of sample size calculations for a single response and a comparison of two responses in a parallel group trial where the outcome is binary. Sample size derivation is given for trials where the objective is to demonstrate: superiority, equivalence, non-inferiority and estimation to a given precision. For each type of trial the null and alternative hypotheses are described and how the impact these have on the sample size calculations. For each type of trial the calculations are highlighted through worked examples. Sample size tables for the different types of trials and worked examples are given to assist in future calculations. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: sample size; power; Type I error; Type II error; superiority trials; equivalence trials; non-inferiority trials; precision; single response; parallel group trials

1. Introduction

An essential step in planning a trial is the calculation of a sample that will give the minimum number required to meet the objectives of the study. We have already given a tutorial for the case where the endpoint is anticipated to have a Normal distribution [1]. This paper extends this work to discuss the situation where the primary endpoint is binary. A review of sample size formulas for the comparison of proportions has been published before [2]; this paper expands and updates that review.

Having as good an estimate as possible of the required sample size is important because studies that are either too small or too large may be judged unethical [3]. For example, a study that is too large could have met the objectives of the trial before the actual study end had been reached, and so some patients may have unnecessarily entered the trial and have been randomised to a therapy that can already be proven to be suboptimal. Conversely a trial that is too small may have little chance of meeting the study objectives, and patients may be entering a trial for no tangible benefit. The general approach to choosing sample size will be described in this article where the primary endpoint can be assumed to be binary and an estimate of the treatment response on at least one of the arms is available. The sections of the paper detail computation of sample sizes appropriate for:

(1) Superiority trials
(2) Equivalence trials
(3) Non-inferiority trials
(4) As good as or better trials
(5) Trials to a given precision

As in our earlier paper [1] a distinction is drawn between trials designed to demonstrate 'superiority' and trials designed to demonstrate 'equivalence' or 'non-inferiority'. We emphasise how differences in the null hypothesis can impact on calculations and in the estimation of the treatment response under the null

University of Sheffield, 30 Regent Court, Regent Street, Sheffield, England, S1 4DA
*Correspondence to: *Steven A. Julious, Reader in Medical Statistics, Medical Statistics Group, ScHARR, University of Sheffield, 30 Regent Court, Regent Street, Sheffield, England, S1 4DA.*
†*E-mail: s.a.julious@sheffield.ac.uk*

and alternative hypotheses [4]. The International Conference on Harmonisation (ICH) guidelines E3 and E9 provide general guidance on selecting the sample size for a clinical trial [5, 6].

Using worked examples, we will also give a brief description of how the calculations can be undertaken in two popular packages PASS 11 [7] and nQuery 7 [8].

The paper is written on the premise that one or two treatments are to be compared in a parallel group trial with a single binary outcome. Each section of the paper will describe the appropriate sample size formulas. Tables are given in each section, which provide sample size estimates using these formulas and worked examples are described that use these tables. Also, within each section some quick approximate formulas are given, which do not require the use of tables for calculations. We assume that the reader is familiar with the concepts of Type I and Type II errors (and power) as discussed in an earlier tutorial article [1] and numerous books including Refs. [4, 9].

2. Single proportion

In studies with a single binary response π_A there are two types of hypotheses that can be investigated depending on whether the objective of the trial is to show that the response is greater than or less than some hypothesised value — as in the null (H_o) and alternative (H_1) hypotheses below.

H_0: The treatment has an effect in terms of the absolute risk being less or equal than some prespecified value ($\pi_A \leqslant \pi_H$).
H_1: The treatment has an effect in terms of the absolute risk being greater than some prespecified value ($\pi_A > \pi_H$).

Alternatively

H_0: The two treatments have equal effect with respect to the absolute risk difference ($\pi_A = \pi_H$).
H_1: The two treatments are different with respect to the absolute risk difference ($\pi_A \neq \pi_H$).

Even if we have a two (or more) arm trial we may still wish to investigate an hypothesis for a single arm. For example the primary endpoint may be based on a continuous scale but we may wish also to show that for a particular adverse event the proportion of events on the investigative treatment arm $\pi_A$can be proved (at a given level of significance) to be less than some *a priori* set clinically important absolute risk, that is, $\pi_A \leqslant \pi_H$.

We will concentrate on the situation of a randomised controlled trial where there is a need to assess a single arm of the trial but without reference to the control — such as in an assessment of adverse events [10]. A way of investigating a single binary response would be to obtain a best estimate of the absolute risk for the investigative treatment and then see if the upper bound (or lower bound depending on the null hypothesis) of the 95% confidence interval for this response excludes the clinically important risk.

2.1. *Confidence interval calculation*

It is worth considering the calculation of confidence intervals for a single binary response before describing the sample size calculations. There are a number of ways of calculating a confidence interval [11]. Here we will concentrate on just two: the Normal approximation approach and the exact method. The Normal approximation is the most common approach for calculating confidence intervals. However, for rare events the Normal approximation may not hold and exact methods may be applied instead.

2.1.1. Normal approximation. Under the Normal approximation the confidence interval for a single proportion is defined as

$$p \pm Z_{1-\alpha/2} se(p), \tag{1}$$

where p is the estimated response from the trial, $se(p) = \sqrt{p(1-p)/n}$, $Z_{1-\alpha/2}$ the $(1-\alpha/2)$% point of the standard Normal distribution and α the level of statistical significance ($\alpha = 0.05$ would give 95% confidence intervals). This method is referred to as the Wald method [11].

2.1.2. Exact confidence intervals. The confidence interval calculations described as 'exact' confidence intervals are also known as Clopper–Pearson confidence intervals [12]. These confidence intervals are calculated by summing each of the tail probabilities from the binomial distribution, given the observed number of cases (k) for the sample size (n). Therefore, defining the individual cell probabilities as

$$\Pr(X = k) = \binom{n}{k} p^{k}(1-p)^{(n-k)}, \tag{2}$$

the lower limit of the confidence interval is calculated as the largest value of p such that the lower tail area of the cumulative distribution is no more than $\alpha/2$. Likewise the upper limit is calculated as the smallest point where the cumulative distribution equals or exceeds $1-\alpha/2$. Formally, the lower point of a confidence interval is defined as the maximum value p_{L} such that

$$\sum_{i=0}^{k} \binom{n}{i} p_{\mathrm{L}}^{i}(1-p_{\mathrm{L}})^{(n-i)} \leqslant \alpha/2, \tag{3}$$

while the upper point is defined as the minimum value p_{U} such that,

$$\sum_{i=0}^{k} \binom{n}{i} p_{\mathrm{U}}^{i}(1-p_{\mathrm{U}})^{(n-i)} \geqslant 1-\alpha/2. \tag{4}$$

An alternative approach to calculate exact confidence intervals would be to use the link between binomial and beta distributions [13, 14]. From this the lower bound is defined as

$$p_{\mathrm{L}} = 1 - BETAINV\left(1-\alpha/2, n-k+1, k\right), \tag{5}$$

and upper as

$$p_{\mathrm{U}} = BETAINV\left(1-\alpha/2, k+1, n-k\right). \tag{6}$$

Here, α is the level of statistical significance ($\alpha = 0.05$ would give 95% confidence intervals), k the number of events observed, n the sample size on the investigative treatment arm and $BETAINV(\bullet)$ refers to the cumulative distribution function of a Beta distribution. The upper and lower bounds calculated from (5) and (6) will provide a range of plausible values that the population proportion is likely to be within. The theoretical rationale behind using the Beta distribution is more complicated than for standard Normal approximation calculations. However, operationally they are easy to calculate and can be calculated in most statistical packages. The $BETAINV(\bullet)$ notation given in this paper is taken from the computer package SAS [15].

Identical confidence intervals can also be obtained using the link between the F distribution and the binomial distribution although with a more complicated nomenclature [11, 13, 14]. This link will not be discussed further.

2.2. *One-tailed or two-tailed?*

The question of whether to calculate one-tailed or two-tailed confidence intervals is not straightforward [16]. It depends on whether we wish to provide an estimate of the plausible range for the true value (two tailed) or a value that you are confident will not be exceeded by the true value (one tailed).

For rare events we are often interested in a one-tailed confidence interval such that a (upper) one-tailed (1–α)% bound is estimated from

$$BETAINV(1-\alpha, k+1, n-k) \tag{7}$$

This one-tailed confidence interval will give an estimate of the proportion for a given number of events k in n subjects, which is unlikely to be exceeded by the true population proportion.

The emphasis in this paper, however, will be on two-tailed confidence interval estimation, that is, using (5) and (6). Because we are only interested in one tail of the 95% confidence interval this would be equivalent to a one-tailed confidence interval but with α set at 2.5%.

2.3. *Sample size calculation*

To calculate the sample size for an anticipated response π_A, which we wish to assess as being less (or greater) than a hypothesised value π_H the following Normal approximation result could be used [17]:

$$n = \frac{\left[Z_{1-\beta}\sqrt{\pi_A(1-\pi_A)} + Z_{1-\alpha/2}\sqrt{\pi_H(1-\pi_H)}\right]^2}{(\pi_A-\pi_H)^2}. \tag{8}$$

This sample size calculation would be consistent with a Normal approximation being used for the confidence interval. Here, α and β are the overall Type I and Type II errors.

An alternative equation is

$$n = \frac{\overline{\pi}(1-\overline{\pi})\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2}{(\pi_A-\pi_H)^2}. \tag{9}$$

where $\overline{\pi} = (\pi_A + \pi_H)/2$. Equation (9) gives similar answers to Equation (8) for $\pi_A < \pi_H$ but gives a slightly larger sample size for $\pi_A > \pi_H$.

Table I gives sample sizes from (8) for various values of $\pi_A > \pi_H$. For $\pi_A < \pi_H$ replace π_A by $1-\pi_A$ and π_H by $1-\pi_H$.

Using a binomial distribution an estimate of the power can be obtained from Equations (10) and (11) [18]

$$\sum_{j=0}^{q}\binom{n}{j}\pi_A^j(1-\pi_A)^{n-j}, \tag{10}$$

where q is the largest integer of k such that

$$\sum_{j=0}^{k}\binom{n}{j}\pi_H^j(1-\pi_H)^{n-j} \leqslant \alpha/2. \tag{11}$$

Table II gives the sample sizes estimate from (10) and (11). As with Table I for $\pi_A < \pi_H$ replace π_A by $1-\pi_A$ and π_H by $1-\pi_H$. It should be noted that here $\pi_A > \pi_H$ and $\pi_A < \pi_H$ do not quite give symmetric results. There were seven instances where the sample size calculated for $\pi_A > \pi_H$ was slightly higher than the sample size estimates 'equivalent' for $\pi_A < \pi_H$. The instances were the discrepancies occurred are: $\pi_A = 0.90$, $\pi_H = 0.95$, $n = 304$, $\pi_A = 0.05$, $\pi_A = 0.05$, $\pi_H = 0.40$, $n = 12$, $\pi_H = 0.30$, $n = 21$, $\pi_A = 0.20$, $\pi_H = 0.25$, $n = 304$, $\pi_A = 0.15$, $\pi_H = 0.25$, $n = 171$, $\pi_A = 0.90$, $\pi_H = 0.25$, $n = 69$ and $\pi_A = 0.95$, $\pi_H = 0.25$, $n = 30$.

Comparing Table I with the equivalent values in Table II we can see that Equation (8) estimates the sample size to be smaller than Equations (10) and (11). This difference in the sample size is due to two reasons. First, Fisher's exact test is more conservative than the asymptotic test and so requires a large sample size for a given significance level and power. Second, because of the discrete nature of the binomial distribution it may be impossible to get an exact Type I and Type II error. One can easily find the actual significance level and power, for a given sample size, that a binomial distribution produces. If these actual values are inserted into (9) we would find results comparable to Tables I and II.

For completeness we include the following result, which gives the sample size where an arcsine transformation (if $y = \sin(x)$ then $x = \arcsin(y)$) is applied to the hypothesised and anticipated responses [19]. The sample sizes estimated from this result are comparable to earlier results and will not be discussed further.

$$n = \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2}{4\left(\arcsin(\sqrt{\pi_A}) - \arcsin(\sqrt{\pi_H})\right)^2}. \tag{12}$$

2.3.1. Worked Example 1 — sample size calculation for a single binary response. An investigator is designing a placebo controlled trial to investigate a new treatment in depression. The sample size for the primary endpoint was calculated to be 525 patients per arm. From experience with other compounds for the same indication the adverse event rate is anticipated to be around 50% in the trial population. The compound under investigation is expected to have a lower adverse event rate of around 40% and

Table I. Sample size calculations for a one-arm trial for a single binary response using the Normal approximation for 90% power and a 95% confidence interval for an alternative hypothesis of $\pi_A > \pi_{\mathrm{H}}$ using (8).

	π_{H}																	
π_{A}	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.10	264	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.15	79	438	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.20	40	122	589	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.25	25	59	158	718	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.30	17	35	74	189	825	—	—	—	—	—	—	—	—	—	—	—	—	—
0.35	12	24	43	87	214	912	—	—	—	—	—	—	—	—	—	—	—	—
0.40	10	17	29	50	97	233	977	—	—	—	—	—	—	—	—	—	—	—
0.45	8	13	20	33	56	105	248	1022	—	—	—	—	—	—	—	—	—	—
0.50	6	10	15	23	36	60	111	257	1045	—	—	—	—	—	—	—	—	—
0.55	5	8	12	17	25	38	62	114	261	1047	—	—	—	—	—	—	—	—
0.60	4	6	9	13	18	26	40	64	115	259	1028	—	—	—	—	—	—	—
0.65	3	5	7	10	14	19	27	40	63	113	252	988	—	—	—	—	—	—
0.70	3	4	6	8	11	14	19	27	40	62	109	240	927	—	—	—	—	—
0.75	2	4	5	6	8	11	14	19	27	38	59	103	222	845	—	—	—	—
0.80	2	3	4	5	7	8	11	14	19	25	36	55	94	200	742	—	—	—
0.85	2	2	3	4	5	7	8	10	13	17	23	33	49	82	171	617	—	—
0.90	1	2	3	3	4	5	6	8	10	12	16	21	28	42	68	137	471	—
0.95	1	2	2	3	3	4	5	6	7	8	10	13	17	23	32	51	96	301

Table II. Sample size calculations for a one-arm trial for a single binary response using a binomial distribution for 90% power and a two-sided significance level of 5% for an alternative hypothesis of $\pi_A > \pi_H$ using Equations (10) and (11). Superscript entries are shown when π_A and π_H are swapped and the results differ.

	π_H																	
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.10	316	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.15	102	492	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.20	55	149	641	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.25	38	75	183	768	—	—	—	—	—	—	—	—	—	—	—	—	—	—
0.30	27	49	90	212	870	—	—	—	—	—	—	—	—	—	—	—	—	—
0.35	18	34	56	103	237	949	—	—	—	—	—	—	—	—	—	—	—	—
0.40	16	23	39	60	112	255	1021	—	—	—	—	—	—	—	—	—	—	—
0.45	14	20	26	42	66	116	266	1066	—	—	—	—	—	—	—	—	—	—
0.50	12	15	20	30	42	70	121	274	1080	—	—	—	—	—	—	—	—	—
0.55	8	13	16	22	31	44	71	125	273	1082	—	—	—	—	—	—	—	—
0.60	7	10	14	18	22	32	46	72	121	275	1059	—	—	—	—	—	—	—
0.65	6	9	11	13	18	24	33	47	69	124	265	1017	—	—	—	—	—	—
0.70	6	8	10	11	15	18	23	30	46	68	117	252	950	—	—	—	—	—
0.75	5	7	9	9	12	13	17	23	32	44	64	111	231	863	—	—	—	—
0.80	5	5	6	8	9	11	14	17	21	31	39	59	99	207	764	—	—	—
0.85	5	5	6	6	7	8	10	11	15	19	27	36	55	87	180	632	—	—
0.90	5	5	5	5	6	8	9	9	13	14	19	21	32	45	73	143	484	—
0.95	5	5	5	5	5	5	7	7	8	9	10	15	18	25	35	51	100	301

the investigator wishes to demonstrate using a 95% confidence interval (equivalent to a Type I error of 2.5%) and 90% power that the event rate is less than 50%.

Since $\pi_A < \pi_H$ we look up 1–π_A and $1 - \pi_H$ and using the Normal approximation result given by Equation (8) the required sample size from Table I for $\pi_H = 0.5$ and $\pi_A = 0.6$ is estimated to be 259 patients. Alternatively using the binomial approach, and results (10) and (11), Table II estimates the sample size to be 275 patients. Because the sample size is less than the sample size of 525 patients being recruited the investigator has sufficient power for the additional objective.

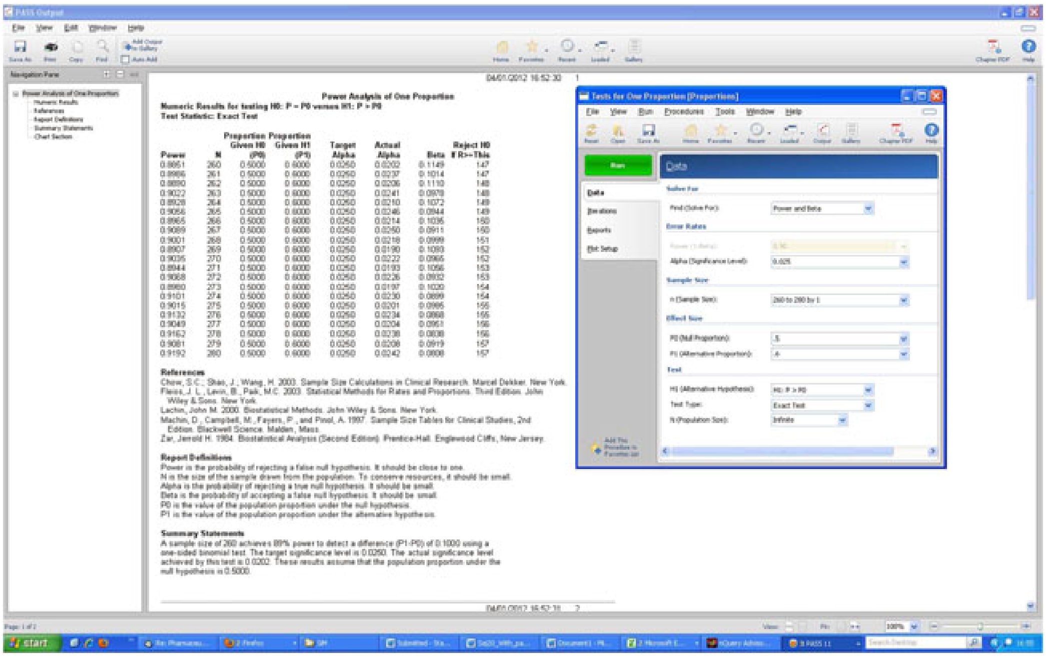
In repeating the calculation in PASS we believe there may be an error in the way PASS estimates sample size for one binomial proportion. To estimate the sample size in PASS you select the menu options *Proportions /One Proportion / Test (Inequality)* and then the icon *Test for One Proportion (Proportions)*. For this problem, the Alternative Hypothesis dialogue can be set to '$p < p0$' and the Type I error as 0.025. PASS can either estimate the sample size (for a given power) or the power (for a given sample size). PASS estimates the sample size to be 263 patients.

The SAS code that seems to mimic the results in PASS is given in Figure 1. This program will iterate until the first integer sample size has greater than 90% power.

There is an interesting issue with this programming approach. Figure 2 gives the power for the study for different sample sizes ranging from 250 to 290 patients for the worked example of $\pi_{\mathrm{A}} = 0.40$ and $\pi_{\mathrm{H}} = 0.50$. We can see from Figure 2 how a power of 90% is obtained for a sample size of 263 patients but now we have less than 90% power for 264 patients! In fact it is not until the sample size is 275 that for both this and the subsequent sample sizes does power exceed 90%. The sample size of 275 patients is what is given in Table II and by Equations (10) and (11).

The reason why the power 'zig–zags' in Figure 2 is due to the discrete nature of the binomial distribution. With additional patients the achievable Type I error may drop, which may mean that the power to achieve that level of significance falls. To estimate the sample sizes given in Table II a program was written so that the iteration will only stop for a given integer sample size if it, and sample sizes up to 10 greater, all had greater than 90% power. Figure 3 gives example SAS code for this calculation.

The equivalent calculation in PASS is given below. Here PASS is run to give the sample size for a range of sample sizes from 260 to 280. PASS now gives a sample size of 274 patients.


PASS Output
File View Edit Window Help
Save As Print Copy Find Add Output to Gallery Auto Add
Home Favorites Recent Loaded Gallery Chapter PDF Help
Navigation Pane
Power Analysis of One Proportion
Numeric Results
References
Report Definitions
Summary Statements
Chart Section
Tests for One Proportion [Proportions]
File View Run Procedures Tools Window Help
Reset Open Save As Home Favorites Recent Loaded Output Gallery Chapter PDF Help
Run
Data
Iterations
Reports
Plot Setup
Page 1 of 2
100%
start

04/01/2012 16:52:30 1

Power Analysis of One Proportion

Numeric Results for testing H0: P = P0 versus H1: P > P0

Test Statistic: Exact Test

		Proportion	Proportion				
		Given H0	Given H1	Target	Actual		Reject H0
Power	N	(P0)	(P1)	Alpha	Alpha	Beta	If R>=This
0.8851	260	0.5000	0.6000	0.0250	0.0202	0.1149	147
0.8986	261	0.5000	0.6000	0.0250	0.0237	0.1014	147
0.8890	262	0.5000	0.6000	0.0250	0.0206	0.1110	148
0.9022	263	0.5000	0.6000	0.0250	0.0241	0.0978	148
0.8928	264	0.5000	0.6000	0.0250	0.0210	0.1072	149
0.9056	265	0.5000	0.6000	0.0250	0.0246	0.0944	149
0.8965	266	0.5000	0.6000	0.0250	0.0214	0.1035	150
0.9089	267	0.5000	0.6000	0.0250	0.0250	0.0911	150
0.9001	268	0.5000	0.6000	0.0250	0.0218	0.0999	151
0.8907	269	0.5000	0.6000	0.0250	0.0190	0.1093	152
0.9035	270	0.5000	0.6000	0.0250	0.0222	0.0965	152
0.8944	271	0.5000	0.6000	0.0250	0.0193	0.1056	153
0.9068	272	0.5000	0.6000	0.0250	0.0226	0.0932	153
0.8980	273	0.5000	0.6000	0.0250	0.0197	0.1020	154
0.9101	274	0.5000	0.6000	0.0250	0.0230	0.0899	154
0.9015	275	0.5000	0.6000	0.0250	0.0201	0.0985	155
0.9132	276	0.5000	0.6000	0.0250	0.0234	0.0868	155
0.9049	277	0.5000	0.6000	0.0250	0.0204	0.0951	156
0.9162	278	0.5000	0.6000	0.0250	0.0238	0.0838	156
0.9081	279	0.5000	0.6000	0.0250	0.0208	0.0919	157
0.9192	280	0.5000	0.6000	0.0250	0.0242	0.0808	157

References

Chow, S.C.; Shao, J.; Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.
Fleiss, J. L., Levin, B., Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.
Lachin, John M. 2000. Biostatistical Methods. John Wiley & Sons. New York.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, Mass.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population. To conserve resources, it should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
P0 is the value of the population proportion under the null hypothesis.
P1 is the value of the population proportion under the alternative hypothesis.

Summary Statements

A sample size of 260 achieves 89% power to detect a difference (P1-P0) of 0.1000 using a one-sided binomial test. The target significance level is 0.0250. The actual significance level achieved by this test is 0.0202. These results assume that the population proportion under the null hypothesis is 0.5000.

04/01/2012 16:52:31 2

Data

Solve For
Find (Solve For): Power and Beta

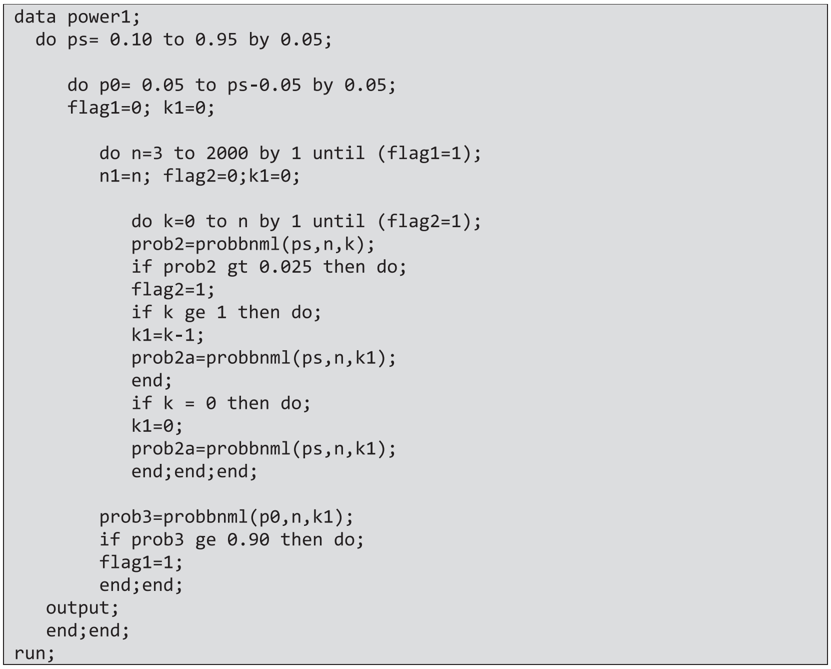
Error Rates
Power (1-Beta): 0.90
Alpha (Significance Level): 0.025

Sample Size
n (Sample Size): 260 to 280 by 1

Effect Size
P0 (Null Proportion): .5
P1 (Alternative Proportion): .6

Test
H1 (Alternative Hypothesis): H1: P > P0
Test Type: Exact Test
N (Population Size): Infinite

PASS also has two other methods giving sample size estimates for two Normal approximation approaches (both with and without a continuity correction). Both approaches are a little different from (9) (and indeed (8)) in that the variance estimate of the treatment effect (under the null or alternative hypothesis) either uses just π_{A} or π_{H} and not both (as in (9) and (8)). For the calculation of the result that uses π_{A} (termed p hat in PASS) the sample size is estimated to be 251 patients.



```
data power1;
  do ps= 0.10 to 0.95 by 0.05;

     do p0= 0.05 to ps-0.05 by 0.05;
     flag1=0; k1=0;

        do n=3 to 2000 by 1 until (flag1=1);
        n1=n; flag2=0;k1=0;

           do k=0 to n by 1 until (flag2=1);
           prob2=probbnml(ps,n,k);
           if prob2 gt 0.025 then do;
           flag2=1;
           if k ge 1 then do;
           k1=k-1;
           prob2a=probbnml(ps,n,k1);
           end;
           if k = 0 then do;
           k1=0;
           prob2a=probbnml(ps,n,k1);
           end;end;end;

        prob3=probbnml(p0,n,k1);
        if prob3 ge 0.90 then do;
        flag1=1;
        end;end;
   output;
   end;end;
run;
```

Figure 1. Example SAS code for calculating sample sizes using a binomial distribution for the alternative hypothesis of $\pi_A > \pi_H$.

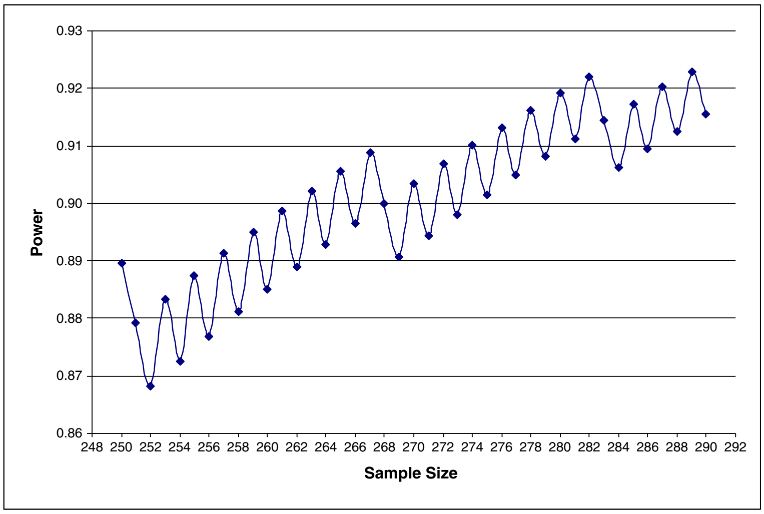
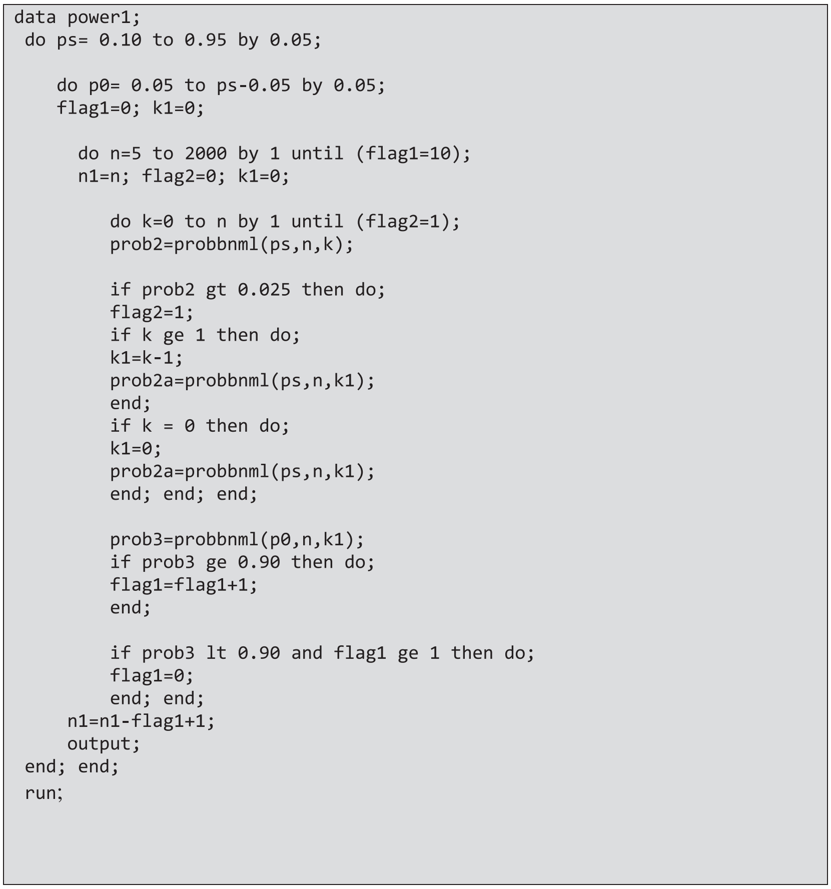


Figure 2. Power for a given sample size for the case $\pi_A = 0.40$ and $\pi_H = 0.50$ where we wish to show that $\pi_A < \pi_H$ with a two-sided 95% confidence interval.

In nQuery to estimate the sample size equivalent to the exact approach given in Equations (10) and (11) the options *Proportions /One /Exact test for a single proportion* need to be selected. nQuery does not give the sample size directly but the power for a given sample size. To save doing many iterations, Equation (9) could be used for an initial sample size. nQuery only gives power to two decimal places in the calculations spreadsheet (actually returns as a percentage equivalent to two decimal places) with more significant digits appearing at the bottom of the window. For the spreadsheet the power is always rounded down. Hence, a power of 89.99 will appear as 89% in the output. nQuery gives a sample size of 274 patients.

To calculate a sample size using a Normal approximation the options *Proportions /One /One sample Chi-squared test* should be ticked. This result gives a sample size of 259 patients, which agrees with nQuery.



```
data power1;
 do ps= 0.10 to 0.95 by 0.05;

    do p0= 0.05 to ps-0.05 by 0.05;
    flag1=0; k1=0;

      do n=5 to 2000 by 1 until (flag1=10);
      n1=n; flag2=0; k1=0;

         do k=0 to n by 1 until (flag2=1);
         prob2=probbnml(ps,n,k);

         if prob2 gt 0.025 then do;
         flag2=1;
         if k ge 1 then do;
         k1=k-1;
         prob2a=probbnml(ps,n,k1);
         end;
         if k = 0 then do;
         k1=0;
         prob2a=probbnml(ps,n,k1);
         end; end; end;

         prob3=probbnml(p0,n,k1);
         if prob3 ge 0.90 then do;
         flag1=flag1+1;
         end;

         if prob3 lt 0.90 and flag1 ge 1 then do;
         flag1=0;
         end; end;
     n1=n1-flag1+1;
     output;
 end; end;
 run;
```

Figure 3. SAS code used to generate Table II for calculating sample sizes using a binomial distribution for the alternative hypothesis of $\pi_A > \pi_H$.

2.4. Sample size calculation revisited — sample size based on feasibility

2.4.1. Precision based approach. As highlighted in Worked Example 1, in clinical trials the primary objective is usually not to estimate a single absolute risk but to compare an investigative treatment with a control for a given objective and endpoint. The sample size would therefore be estimated from the primary endpoint and hence 'fixed' with respect to the objective for the single absolute risk. In this context therefore the objective may not be to prove a risk is less than some bound but to quantify the likely range of values that the risk could plausibly be — through a confidence interval.

In Section 3.5 precision-based trials are described where the objective is to quantify the risk difference against control.

For a single risk the precision of the trial can be estimated from

$$w = \frac{Z_{1-\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \tag{13}$$

where w here is defined as half the width for a confidence interval. Here, it is assumed that both n and p are known. To estimate a sample size to have a required precision, w, about p then the following result could be used:

$$n = \frac{Z_{1-\alpha/2}^2 p(1-p)}{w^2}. \tag{14}$$

If exact confidence intervals are being used, then we can estimate the precision for a trial from

$$w = (BETAINV\,(1-\alpha/2, k+1, n-k) + BETAINV\,(1-\alpha/2, n-k+1, k) - 1)\,/2, \tag{15}$$

where k is estimated from $k = pn$. To estimate the sample size, we can iterate on n until we get a sample size with the requisite precision for a given p.

Table III. Probablilities of observing a given number of adverse events or more (k) for given anticipated risks for a sample size of 100 patients.

	Risk of an event					
k	0.0500	0.0300	0.0100	0.0050	0.0010	0.0001
1	0.9941	0.9524	0.6340	0.3942	0.0952	0.0100
2	0.9629	0.8054	0.2642	0.0898	0.0046	< 0.0001
3	0.8817	0.5802	0.0794	0.0141	0.0002	—
4	0.7422	0.3528	0.0184	0.0017	< 0.0001	—
5	0.5640	0.1821	0.0034	0.0002	—	—
6	0.3840	0.0808	0.0005	< 0.0001	—	—
7	0.2340	0.0312	0.0001	—	—	—
8	0.1280	0.0106	< 0.0001	—	—	—
9	0.0631	0.0032	—	—	—	—
10	0.0282	0.0009	—	—	—	—

2.4.2. Probability of seeing an event. In the context of clinical trials, the results described above may not be readily applicable when we wish to quantify a risk particularly if this risk is quite rare, such as with an adverse event. A more appropriate calculation quantifies the probability of seeing the event for the finite (and fixed) sample size in the trial.

Hence, if the risk for a particular adverse event is p then the probability that k or more adverse events will be observed with n subjects can be calculated from

$$p_k = 1 - \sum_{x=0}^{k-1} \binom{n}{x} p^x (1-p)^{n-x}. \tag{16}$$

Worked Example 2 — calculating a probability of observing an adverse event.

A Phase II trial has been designed where the number of patients per arm is 100. For the investigative treatment, a number of adverse events are being monitored with different anticipated risks. Table III gives the probability of observing various numbers of adverse events for different anticipated population risks.

From Table III we can see that for a risk of an adverse event of 1/1000 we would have less than 10% chance of observing at least one adverse event. Also, for a risk of 1/10,000 we would have only a 1% chance of seeing at least one adverse event.

We recommend that a table such as Table III be calculated for all planned clinical trials.

In Table III could be used to put the results into some context if no adverse events are observed. This could be conducted in context also with the '3 over n' ($3/n$) rule [10]. The $3/n$ rule gives the approximate upper tail of a one-sided 95% confidence interval when zero events are observed and is derived using the Poisson approximation to Equation (16).

Suppose there are no observed instances of a particular adverse event in the trial of 100 subjects we are describing. Suppose in the protocol we had stated that for the adverse event we anticipated the population risk to be $1/2000 = 0.005$ and thus that *a priori* we would anticipate that there was a probability of 0.39 of observing at least one adverse event. We could highlight this probability when discussing the result. Also, we could state that, based on the observed trial data, we can rule out a risk of 3/100 $(3/n) = 0.03$ or 3% or greater.

3. Parallel group trials

3.1. Superiority trials

With a superiority trial the objective is to determine whether there is evidence of a statistical difference in the comparison of interest between the regimens with reference to the null hypothesis that the regimens are the same. The null (H_0) and alternative (H_1) hypotheses may take the form:

H_0: The two treatments are not different ($\pi_A = \pi_B$).
H_1: The two treatments are different ($\pi_A \neq \pi_B$) that is, either A is superior to B or B is superior to A.

For a two-sided superiority trial there are two chances of rejecting the null hypothesis and thus making a Type I error. The null hypothesis can be rejected if $p_A > p_B$ or if $p_A < p_B$ by a statistically significant

Table IV. Summary table for a clinical trial with a binary outcome.

	Outcome		
Treatment	1	0	Sample size
A	p_A	$1 - p_A$	n_A
B	p_B	$1 - p_B$	n_B
Overall response	$\overline{p} = (n_A p_A + n_B p_B)/(n_A + n_B)$	$1 - \overline{p}$	$n = n_A + n_B$

amount. Because there are two chances of rejecting the null hypothesis the statistical test is referred to as a two-tailed test with each tail allocated an equal amount of the Type I error (of 2.5%). The sum of these tails adds up to the overall Type I error rate of 5%. Thus, the null hypothesis can be rejected if the test of $\pi_A > \pi_B$ is statistically significant at the 2.5% level of significance or the test of $\pi_A < \pi_B$ is statistically significant at the 2.5% level.

The purpose of the sample size calculation is hence to provide sufficient power to reject H_0 when in fact some alternative hypothesis is true.

3.1.1. Summarising clinical trials with binary data. For a clinical trial where the primary outcome is a binary response the notation is given in Table IV where p_A and p_B are the responses anticipated on treatment A and B, respectively, $\overline{p}$ is the average response across treatments, n_A and n_B are the sample sizes in each treatment group and n is the total sample size.

The absolute risk reduction is probably the simplest way of summarising binary data, which is $p_A - p_B$, and this is the scale that we will focus on.

One drawback of working with the absolute risk difference is that it is bounded by (–1, 1). This bounding can adversely affect inference — especially when both responses are near one of the bounds.

3.1.2. Sample sizes for a superiority trial. For the special case of equally sized arms in the trial the sample size is

$$n_A = \frac{\left(Z_{1-\alpha/2}\sqrt{2\overline{\pi}(1-\overline{\pi})} + Z_{1-\beta}\sqrt{\pi_A(1-\pi_A) + \pi_B(1-\pi_B)}\right)^2}{(\pi_A - \pi_B)^2} \tag{17}$$

(where $\overline{\pi} = (\pi_A + \pi_B)/2$ the average response).

Since the expressions under the square roots are relatively stable to changes in the π's, this is often simplified to [4, 20, 21]

$$n_A = \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 (\pi_A(1-\pi_A) + \pi_B(1-\pi_B))}{(\pi_A - \pi_B)^2}. \tag{18}$$

The result (18) gives the maximum sample size for the case where $\overline{\pi} = 0.5$ [4]. From this fact and within this range for the average response a quick estimate of the sample size, for 90% power and two-sided significance level of 5%, can be obtained from the following result [4]:

$$n_A = \frac{5.25}{(\pi_A - \pi_B)^2}. \tag{19}$$

For 80% power and two-sided significance level of 5% the sample size can be estimated from [4]

$$n_A = \frac{4}{(\pi_A - \pi_B)^2}. \tag{20}$$

Both of these results will provide conservative 'maximum' estimates of the sample size.

For these sample size calculations we have assumed equal allocation to treatment. For fixed allocation to treatment there are extensions to these results [21], and for random allocation there are alternative results [22]. In addition we have assumed there will be just a single endpoint in the trial. For calculations with multiple endpoints there are alternative calculations [23–25].

Sample sizes for selected values of π_A and π_B using Equation (17) are given in Table V.

Table V. Sample size estimates using result (17) for one arm of a parallel group trial for various expected outcome responses for a given treatment (π_A) and comparator (π_B) for a two-sided Type I error rate of 5% and 90% power.

	π_B								
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	582	—	—	—	—	—	—	—	—
0.15	188	918	—	—	—	—	—	—	—
0.20	101	266	1212	—	—	—	—	—	—
0.25	65	133	335	1464	—	—	—	—	—
0.30	47	82	161	392	1674	—	—	—	—
0.35	36	57	97	185	440	1842	—	—	—
0.40	28	42	65	109	203	477	1969	—	—
0.45	23	33	47	72	118	217	503	2053	—
0.50	19	26	36	52	77	124	227	519	2095
0.55	16	21	28	39	54	81	128	231	524
0.60	14	17	23	30	40	56	82	130	231
0.65	12	15	19	24	31	41	57	82	128
0.70	10	12	15	19	24	31	41	56	81
0.75	8	10	13	16	19	24	31	40	54
0.80	7	9	11	13	16	19	24	30	39
0.85	6	7	9	11	13	15	19	23	28
0.90	5	6	7	9	10	12	15	17	21
0.95	4	5	6	7	8	10	12	14	16

If we intend to use a continuity corrected chi-squared test in the analysis then (17) and (18) could be used to estimate initial values of the sample size, which are then increased to account for the conservative nature of this test using the following result [21].

$$n_{cc} = \frac{n_A}{4}\left[1+\sqrt{1+\frac{4}{n_A(\pi_A-\pi_B)}}\right]^2 \qquad (21)$$

Table VI gives estimates of the sample size using Equation (21) with (17).

Table VI. Sample size estimates using result (17) with a continuity correction (21) for one arm of a parallel group trial for various expected outcome responses for a given treatment (π_A) and comparator (π_B) for a two-sided Type I error rate of 5% and 90% power.

	π_B								
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	622	—	—	—	—	—	—	—	—
0.15	208	958	—	—	—	—	—	—	—
0.20	114	286	1252	—	—	—	—	—	—
0.25	75	147	355	1504	—	—	—	—	—
0.30	55	92	175	412	1714	—	—	—	—
0.35	43	65	107	199	460	1882	—	—	—
0.40	34	49	73	119	217	497	2009	—	—
0.45	28	39	54	80	128	231	523	2093	—
0.50	24	31	42	59	85	134	241	539	2135
0.55	20	26	33	45	61	89	138	245	544
0.60	18	21	28	35	46	63	90	140	245
0.65	16	19	23	29	36	47	64	90	138
0.70	13	16	19	23	29	36	47	63	89
0.75	11	13	17	20	23	29	36	46	61
0.80	10	12	14	17	20	23	29	35	45
0.85	9	10	12	14	17	19	23	28	33
0.90	8	9	10	12	13	16	19	21	26
0.95	7	8	9	10	11	13	16	18	20

If the final analysis is to be a Fisher's exact test then the sample size calculation is not so straightforward. The sample size is calculated in two stages. Conditional on the number of events observed k_A in n_A subjects on treatment A and k_B events in n_B subjects on treatment B such that $k = k_A + k_B$ and $n = n_A + n_B$, we can use a hypergeometric distribution to find the probability of a number of events $k_i < n_A$ as

$$P_{k_i} = P(k_i|k,n,n_A) = \frac{\binom{n_A}{k_i}\binom{n_B}{k-k_i}}{\binom{n}{k}}. \tag{22}$$

The P-value is defined as the sum of all the P_{k_i}, which are $\leqslant P_{k_A}$, that is,

$$F(k_A|n,k,n_A) = \sum_{k_i=0}^{k_A} \frac{\binom{n_A}{k_i}\binom{n_B}{k-k_i}}{\binom{n}{k}}. \tag{23}$$

For the subset of tables where we reject the null hypothesis from (22) we can estimate the power under the alternative hypothesis, in Equation (24).

$$Power = \sum\sum \binom{n_A}{k_A}\binom{n_B}{k_B} \pi_A^{k_A}(1-\pi_A)^{n_A-k_A}\pi_B^{k_B}(1-\pi_B)^{n_B-k_B} \quad . \tag{24}$$

Thus, for a given k_A, k_B, n_A and n_B we can estimate the power. Hence, through iteration we can estimate the sample size for a given π_A and π_B for a given nominal power. As for a single binary response discussed earlier in the paper, we need to iterate beyond the sample size achieved when first a power of 90% is reached. For the programming in this paper the program stopped once a sample size had a power greater than 90% and all the sample sizes up to at least 10 subjects more also all had power greater than 90%.

Table VII gives sample sizes for 90% power for a one-tailed Type I error of 2.5%, which will be taken to be the same as for a two-tailed Type I error of 5%.

Table VII. Sample size estimates for one arm of a parallel group trial for various expected outcome responses for a given treatment (π_A) and comparator (π_B) for a one-sided Type I error rate of 2.5% and 90% power assuming Fisher's exact test is the final analysis.

	π_B								
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	605	—	—	—	—	—	—	—	—
0.15	188	965	—	—	—	—	—	—	—
0.20	108	285	1264	—	—	—	—	—	—
0.25	62	142	362	1502	—	—	—	—	—
0.30	49	89	175	415	1731	—	—	—	—
0.35	33	65	108	202	468	1876	—	—	—
0.40	30	47	72	118	219	502	2029	—	—
0.45	28	37	55	81	133	235	526	2075	—
0.50	19	29	43	59	87	133	243	550	2151
0.55	18	25	32	47	62	87	125	228	520
0.60	16	23	26	37	48	67	91	126	228
0.65	13	17	23	29	38	47	67	94	125
0.70	10	13	16	23	28	38	47	65	91
0.75	9	12	17	21	24	29	36	47	58
0.80	9	11	12	17	20	22	29	33	40
0.85	8	10	12	15	15	20	22	27	29
0.90	8	7	10	10	12	15	15	20	20
0.95	4	7	7	10	10	10	10	12	15

It is interesting to compare Table VI with Table VII. The two tables are reasonably comparable and so if a Fisher's exact test is to be considered for the final analysis it may be worth estimating the sample size using the more straightforward approach of the continuity corrected sample size calculation.

The programming for Table VII is quite computer intensive. A quick estimate of the sample size for Fisher's exact test can be obtained from a simple Normal approximation. If, in a study, we actually observed the predicted effect size, with the required sample size at significance level α and power 1–β, then the observed test statistic is simply $z_{1-\alpha} + z_{1-\beta}$. For α of 0.05 and β of 0.10 the one-sided P-value would actually be 0.00059. Thus, a quick method of obtaining the correct sample size is to perform Fisher's exact test on the given proportions with increasing sample size until a one-sided P-value of 0.00059 is obtained. The result of this procedure is given in Table VIII. This quick method is quite useful generally and deserves to be better known.

The results in Tables VII and VIII are reasonably close. The advantage of the approach in Table VIII is that it is quite easy to program and the SAS code is given in Figure 4.

Table VIII. Sample size estimates for one arm of a parallel group trial for various expected outcome responses for a given treatment (π_A) and comparator (π_B) for a one-sided P-value of 0.059% assuming Fisher's exact test is the final analysis.

	π_B								
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	615	—	—	—	—	—	—	—	—
0.15	204	977	—	—	—	—	—	—	—
0.20	113	298	1258	—	—	—	—	—	—
0.25	74	150	358	1514	—	—	—	—	—
0.30	55	92	179	429	1739	—	—	—	—
0.35	42	68	108	205	468	1896	—	—	—
0.40	37	49	74	124	227	507	2017	—	—
0.45	26	39	55	84	135	237	526	2095	—
0.50	23	31	45	59	87	137	243	545	2131
0.55	21	29	38	49	67	96	143	250	560
0.60	18	21	31	36	48	66	95	145	251
0.65	17	20	26	30	40	47	66	95	143
0.70	15	18	20	25	31	40	51	65	91
0.75	13	16	18	21	25	30	40	48	61
0.80	12	12	15	19	21	26	31	36	49
0.85	12	12	14	15	19	20	26	28	38
0.90	9	10	12	12	15	18	20	21	29
0.95	7	9	12	12	15	16	17	18	21

```
data power;
    do pa=0.10 to 0.95 by 0.05;
      do pb=0.05 to pa-0.45 by 0.05;
      flag=0;
      p=round(10.5*(pa*(1-pa)+pb*(1-pb))/((pa-pb)*(pa-pb)))-1+3;
        do n=p to 10000 by 1 until (flag=10);
        ka=round(pa*n);
        kb=round(pb*n);
        m=ka+kb;
        prob=probhypr(2*n,m,n,kb);
        if prob lt 0.00059 then do;
        flag=flag+1;
        end;
        if prob ge 0.00059 and flag ge 1 then do;
        flag=0;
        end;
        end;
        n=n-flag+1;
        output;
        end;end;run;
```

Figure 4. SAS code used to generate Table VIII for calculating sample sizes using only the P-value.

If we planned to use a mid-P P-value with Fisher's exact test, then Table IX gives sample sizes for 90% power for a one-tailed Type I error of 2.5%. This is calculated by amending Equation (22) to become Equation (25)

$$F(k_A|n,k,n_A) = \sum_{k_i=0}^{k_A-1} \frac{\binom{n_A}{k_i}\binom{n_B}{k-k_i}}{\binom{n}{k}} + 0.5\left(\sum_{k_i=0}^{k_A} \frac{\binom{n_A}{k_i}\binom{n_B}{k-k_i}}{\binom{n}{k}} - \sum_{k_i=0}^{k_A-1} \frac{\binom{n_A}{k_i}\binom{n_B}{k-k_i}}{\binom{n}{k}}\right) \quad (25)$$

Or alternatively

$$F(k_A|n,k,n_A) = \sum_{k_i=0}^{k_A-1} \frac{\binom{n_A}{k_i}\binom{n_B}{k-k_i}}{\binom{n}{k}} + 0.5\frac{\binom{n_A}{k_A}\binom{n_B}{k_B}}{\binom{n}{k}}. \quad (26)$$

The one-sided mid-P P-value is defined using (23) as the sum of the P_{k_i}, which are less than P_{k_A-1} plus half the value of P_{k_A} from (23).

Comparing Table IX to Table VII we see there are bigger (in absolute terms) differences in the sample size estimates for the smallest effect sizes. The sample sizes estimated in Table IX are closer to those of Table V.

We can repeat the quick method used in Table VIII for a mid-P value by replacing the line

$$prob = probhypr(2^*n, m, n, kb);$$

Table IX. Sample size estimates for one arm of a parallel group trial for various expected outcome responses for a given treatment (π_A) and comparator (π_B) for a one-sided Type I error rate of 2.5% and 90% power assuming mid-P Fisher's exact test is the final analysis.

	π_B								
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	577	—	—	—	—	—	—	—	—
0.15	188	917	—	—	—	—	—	—	—
0.20	96	268	1213	—	—	—	—	—	—
0.25	61	134	341	1478	—	—	—	—	—
0.30	45	82	169	405	1682	—	—	—	—
0.35	31	60	102	193	455	1837	—	—	—
0.40	29	45	67	109	204	489	1979	—	—
0.45	19	33	50	76	122	225	506	2055	—
0.50	17	25	33	51	81	127	237	526	2109
0.55	16	23	29	37	56	83	125	228	520
0.60	15	16	23	30	45	59	83	126	228
0.65	10	14	21	26	29	44	59	88	125
0.70	10	11	16	21	27	35	44	57	81
0.75	9	10	13	19	21	26	33	42	55
0.80	9	10	11	14	19	21	26	33	37
0.85	8	9	9	11	14	16	19	21	26
0.90	8	6	9	9	9	11	14	16	19
0.95	4	6	6	6	9	9	9	9	11

Table X. Sample size estimates for one arm of a parallel group trial for various expected outcome responses for a given treatment (π_A) and comparator (π_B) for a mid-P P-value 0.059% assuming Fisher's exact test is the final analysis.

	π_B								
π_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	580	—	—	—	—	—	—	—	—
0.15	186	916	—	—	—	—	—	—	—
0.20	99	265	1210	—	—	—	—	—	—
0.25	64	132	333	1462	—	—	—	—	—
0.30	45	81	160	391	1672	—	—	—	—
0.35	34	55	95	183	438	1840	—	—	—
0.40	27	41	64	107	202	475	1966	—	—
0.45	21	31	46	70	116	216	501	2050	—
0.50	17	24	34	50	76	123	225	517	2092
0.55	14	20	27	37	53	80	127	230	522
0.60	13	16	21	28	39	55	81	128	230
0.65	10	13	18	22	29	40	57	81	127
0.70	8	11	14	18	23	30	40	55	80
0.75	7	9	12	14	18	23	29	39	53
0.80	6	9	9	14	15	18	22	28	37
0.85	5	7	7	9	12	14	18	21	27
0.90	4	7	7	9	9	11	13	16	20
0.95	3	4	5	6	7	8	10	13	14

with

$$prob = probhypr(2^{*}n, m, n, kb-1) + 0.5^{*}(probhypr(2^{*}n, m, n, kb) - probhypr(2^{*}n, m, n, kb-1));$$

in Figure 4. The sample sizes are given in Table X and are reasonably close to those of Table IX.

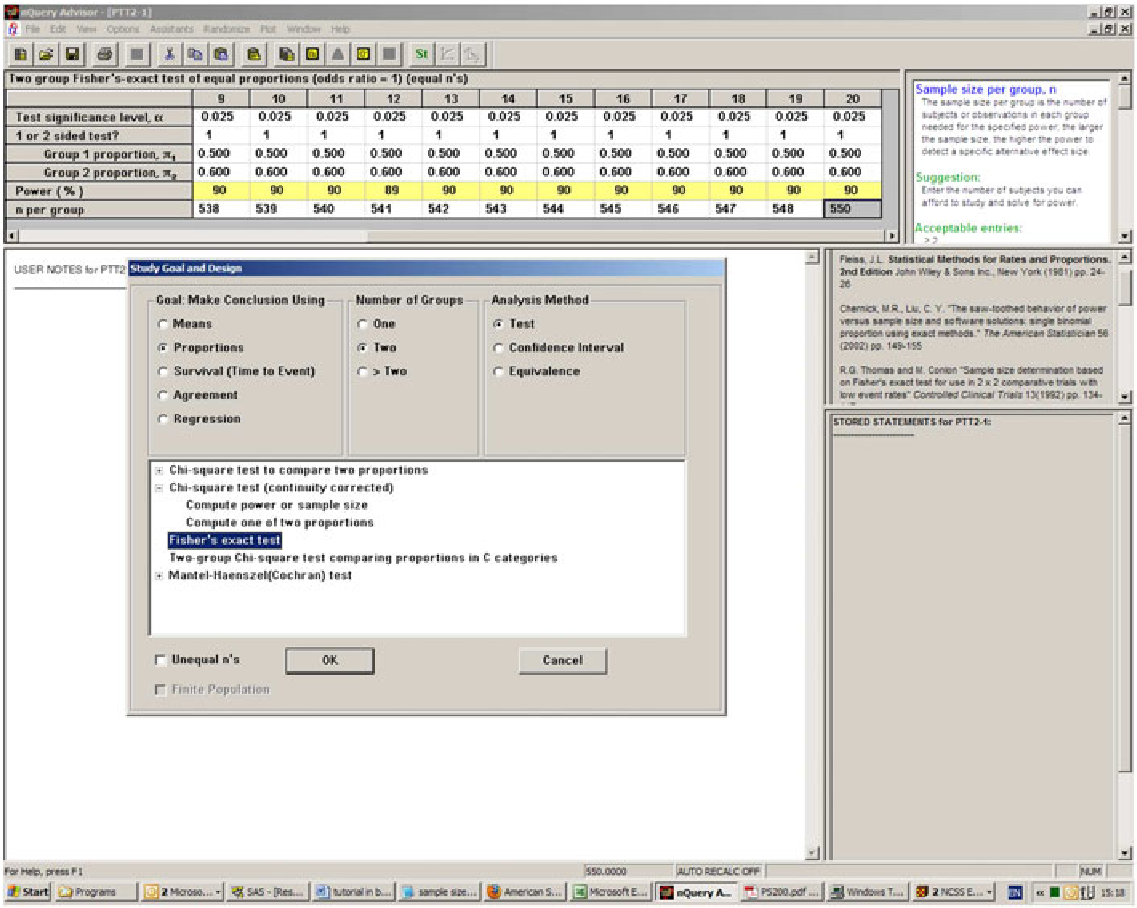
3.1.3. Worked Example 3 — sample size calculation for a parallel group superiority trial with binary response. An investigator wishes to design a placebo controlled trial to investigate a new treatment for migraine. The absolute risk of migraine on placebo over the trial period is anticipated to be 50% and it would be clinically worthwhile using the drug if the risk was reduced on the new treatment to 40%. This is a treatment effect of an absolute risk reduction of 10%. The investigator wished to design the study to have 90% power and a two-sided significance level of 5%.

The sample sizes using the different methods are given in Table XI.

To repeat the calculations in nQuery you need to select File/New and then under 'Make Conclusions Using' tick 'Proportions'; under 'Number of Groups' tick 'Two' and under 'Analysis Method' tick 'Test'. nQuery will then give you three options 'Chi-squared test', which seems to be equivalent to calculation from (17); 'Chi-squared test (continuity corrected)', which seems to be based on Equation (17) with (21) and Fisher's exact test. The calculations for Fisher's exact test are given below. nQuery does not give the sample size for this calculation but rather the power for a given sample size. You then need to iterate the required sample size by hand —remembering not to stop just because a sample size gives a power of 90%. For Fisher's exact test nQuery gave a sample size of 542 patients per arm. A comparison of the results from nQuery with PASS and those in the paper are given in Table XI.

Table XI. Comparison of results in current paper with nQuery and PASS.

	Current paper	nQuery	PASS
Normal approximation from (18)	515	N/A	515
Normal approximation from (17)	519	519	519
Continuity correction from (17) and (21)	539	538	538
Fisher's exact test	550	542	533
Fisher's exact test mid-P	526	N/A	N/A


nQuery Advisor - [PTT2-1]
File Edit View Options Assistants Randomize Plot Window Help
USER NOTES for PTT2
For Help, press F1
550.0000
AUTO RECALC OFF
NUM

Two group Fisher's-exact test of equal proportions (odds ratio = 1) (equal n's)

	9	10	11	12	13	14	15	16	17	18	19	20
Test significance level, α	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025
1 or 2 sided test?	1	1	1	1	1	1	1	1	1	1	1	1
Group 1 proportion, π_1	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Group 2 proportion, π_2	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600
Power (%)	90	90	90	89	90	90	90	90	90	90	90	90
n per group	538	539	540	541	542	543	544	545	546	547	548	550

Sample size per group, n
The sample size per group is the number of subjects or observations in each group needed for the specified power; the larger the sample size, the higher the power to detect a specific alternative effect size.

Suggestion:
Enter the number of subjects you can afford to study and solve for power.

Acceptable entries:
> 2

Fleiss, J.L. **Statistical Methods for Rates and Proportions. 2nd Edition** John Wiley & Sons Inc., New York (1981) pp. 24-26

Chernick, M.R., Liu, C. Y. "The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods." *The American Statistician* 56 (2002) pp. 149-155

R.G. Thomas and M. Conlon "Sample size determination based on Fisher's exact test for use in 2 x 2 comparative trials with low event rates" *Controlled Clinical Trials* 13(1992) pp. 134-

STORED STATEMENTS for PTT2-1:

Study Goal and Design

Goal: Make Conclusion Using
○ Means
● Proportions
○ Survival (Time to Event)
○ Agreement
○ Regression

Number of Groups
○ One
● Two
○ > Two

Analysis Method
● Test
○ Confidence Interval
○ Equivalence

+ Chi-square test to compare two proportions
- Chi-square test (continuity corrected)
  Compute power or sample size
  Compute one of two proportions
 Fisher's exact test
 Two-group Chi-square test comparing proportions in C categories
+ Mantel-Haenszel(Cochran) test

□ Unequal n's OK Cancel

□ Finite Population

In PASS, to calculate the sample size, you need to select 'Proportions' and the 'Two Groups: Independent' and finally 'Inequality (Proportions)'. You can then drop down in the dialogue box 'Test for' to calculate sample sizes for 'Z-test unpooled' (equivalent to (18)); 'Z-test pooled' (equivalent to (17)); 'Z-test cc pooled' (equivalent to (17) and (21)) and 'Fisher's Exact Test'. For Fisher's exact test the calculation is only performed as default if the sample sizes in each arm are both less than 100. If this is not the case, then the continuity corrected calculations (equivalent to Equations (18) and (21)) are undertaken. To change the default click on options and under 'Exact Test Options' reset the 'Maximum N1 or N2 for Exact Calculations', for example to 10,000. For Fisher's exact test PASS gives a sample size of 533 patients per arm.

There is a similar issue with PASS for two-arm trials as for a single-arm trial highlighted in Figure 2. Figure 5 highlights how PASS crosses the power boundary of 90% for a sample size of 533 before dropping below it again and recrossing at 542 patients per arm.

Neither PASS or nQuery gives sample size estimates for mid-P P-values.

Our results are very slightly larger than those of nQuery for this worked example using Fisher's exact test. For the continuity corrected sample size estimation PASS and nQuery give a sample size one less than the results in the paper. We suspect this may be due to the steps used for sample size calculation. For the Normal approximation using Equation (18) both PASS and nQuery and this paper estimate the sample size to be 519 patients per arm. In actuality this was 518.04 rounded up to 519. If 519 is then used in (21) the sample size is estimated to be 539 patients. If 518.04 is used instead the sample size is 538 patients per arm.

3.1.3. Discussion of the sample size calculations. There is a maxim that you should analyse your study as you have designed it. With sample size calculations it is the opposite way — your design should reflect your planned analysis. Hence, if the plan is to undertake a chi-squared test for the primary analysis, then a sample size calculation should reflect this. Thus, for both a single-arm trial and a two-arm trial depending on the assumptions for the analysis, the planned statistical test should be considered [26, 27].

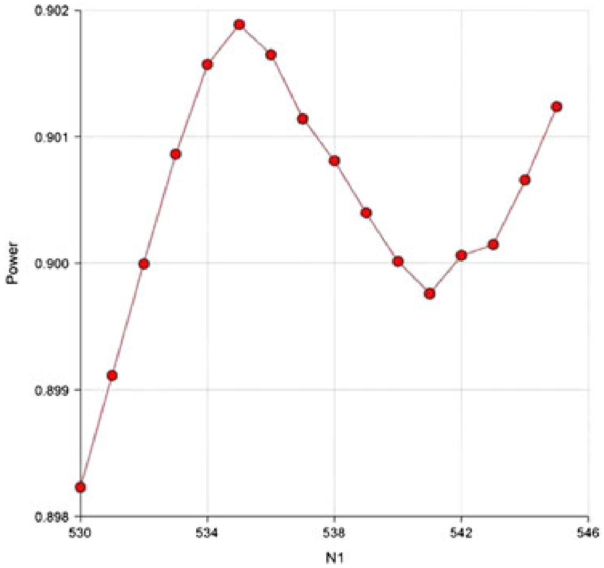


Figure 5. Power for a given sample size for the case $\pi_A = 0.40$ and $\pi_B = 0.50$ for a Fishers exact test for a one-sided Type I error rate of 2.5% from PASS.

We would recommend that the simple asymptotic approaches described here should be used for most sample size calculations. This does not preclude other approaches being used (including maybe simulations) to investigate the sensitivity of the initial calculations.

3.2. *Non-inferiority trials*

In the initial investigation of a new therapeutic intervention for a particular disease, randomised trials are conducted against either placebo or a 'treated as usual' control group. However, when the existing therapy has been established as effective, it may then be no longer ethical to undertake randomised trials where the control therapy is placebo. Instead active-controlled trials are conducted where the new treatment is compared with an established treatment with the objective of demonstrating that new treatment is non-inferior to this established treatment.

For certain trials therefore the objective is not to demonstrate that two treatments are different but rather to demonstrate that a given treatment is clinically not inferior compared with another. The null (H_0) and alternative (H_1) hypotheses for non-inferiority trials may take the form:

H_0: A given treatment is inferior with respect to the absolute risk of a response.
H_1: A given treatment is non-inferior with respect to the absolute risk of a response.

A non-inferiority study is usually planned therefore to detect if the effect of the investigative treatment is not much worse than the control treatment defined by a non-inferiority margin, d. An assessment of non-inferiority of a new treatment is usually performed by comparing the lower tail of 95% confidence interval with the non-inferiority margin to rule out the inferiority of a new treatment. The threshold setting of d is not straightforward and is defined as the largest difference that is clinically acceptable such that a larger difference than this would matter in clinical practice [28]; a clinical judgement. This difference also cannot be 'greater than the smallest effect size that the active (control) drug would be reliably expected to have compared with placebo in the setting of the planned trial' [29]; a statistical assessment. Often the margin is defined as some fraction of the active control effect (over placebo) to be retained and the control effect is estimated from historical trials as a statistical margin. Jones *et al.* [30] recommend that the choice of limit be set at half the expected clinically meaningful difference between the active control and placebo as a clinical margin. For a binary outcome, the active control effect may be expressed as the difference or difference in the logarithms in the event rates, or the difference in log-odds of the event of interest. Generally, the definition of an acceptable level of non-inferiority is made with reference to some retrospective superiority comparison to placebo [31–34]. In this context we layout the assumptions in a two-arm non-inferiority trial and the issues with the non-inferiority margin [1, 30–35]. There are regulatory guidelines on setting the non-inferiority margin [34, 36].

Thus, the two hypotheses become:

H_0: $\pi_\mathrm{A} - \pi_\mathrm{B} \leqslant -d$.
H_1: $\pi_\mathrm{A} - \pi_\mathrm{B} > -d$.

In the context of non-inferiority trials $-d$ is known as the non-inferiority limit.

To conclude non-inferiority, we need to reject the null hypothesis. Thus, non-inferiority trials reduce to a simple one-sided hypothesis test. In practice, this is operationally the same as constructing a $(1-2\alpha)100\%$ confidence interval and concluding non-inferiority provided that the lower end of this confidence interval is greater than $-d$.

To analyse a non-inferiority trial, the following ABC should be considered [37, 38]:

1. The **A**ssay sensitivity of the active control in both the placebo controlled trials and in the active controlled non-inferiority trial exists.
2. **B**ias is minimised through steps such as ensuring that the patient population and the primary efficacy endpoint are essentially the same for the placebo-controlled trial and the active-controlled trial.
3. **C**onstancy assumption of the effect of the common comparator. For two trials in sequence, Trial 1 and Trial 2, the control effect of Treatment B versus Placebo in Trial 1 is assumed to be the same as the control effect of Treatment B versus 'Placebo' in Trial 2.

In addition, to demonstrate that there is no clinically meaningful inferiority of the investigative treatment compared with the active control comparator, non-inferiority studies often entail an indirect cross-trial assessment. The indirect inference is that through comparing the investigative treatment to the control treatment, whether a new treatment preserves a fraction of the control effect or is superior to the 'placebo' not concurrently studied.

This is an issue, however, in that the estimate of effect over placebo in Trial 1 may possibly be overestimated for comparison in Trial 2 because of the placebo responses improving over time, that is, placebo 'creep'. However, the lack of constancy of control effect prescribed by the placebo 'creep' cannot be formally tested [30–35, 39, 40], although an educated assessment of constancy violation may help [41].

To ensure the choice of margin and hence to ensure the study is not biased, the following factors are critical in defining the non-inferiority margin:

(1) How should the heterogeneity of the control effect and its variability across completed placebo-controlled trials, relative to Trial 1, be incorporated?
(2) Should differential weight be given to the response from the most recent studies and/or from the studies with smaller effects?
(3) What should be the preservation fraction be to account for the placebo 'creep'?

From a public health perspective, what we wish to do is to protect the efficacy that has been established with the standard therapy when undertaking non-inferiority trials. This is as it is described for vaccination trials for example [42].

Non-inferiority studies are often thought of as trials where there is a need to make an indirect comparison with placebo using the active control in the current trial. Indirect comparisons are undertaken when a comparison is made between two regimens where the regimens have usually never been given concurrently in any controlled trial investigating the same general patient population. To make comparisons of the regimens of interest, common controls from the trials undertaken for these regimens are used. For example consider Scenario 1 where two trials were conducted with the following regimens randomised:

Trial 1: Placebo and Treatment A,
Trial 2: Placebo and Treatment B.

We could use the fact that both regimens have had a trial where they were compared with placebo to make comparisons between treatments A to B in the same patient population and the same primary efficacy endpoint studied.

Now consider Scenario 2 where Trial 1 and Trial 2 are conducted in sequence with the following setup.

Trial 1: Placebo and Treatment A,
Trial 2: Treatment A and Treatment B.

Treatment A should have been shown to be effective in Trial 1 (a placebo-controlled trial) to launch Trial 2 (an active-controlled trial). In some disease areas, when an approved agent becomes the standard of care it may no longer be ethical to conduct a placebo-controlled trial. Thus, because of ethical constraints, Trial 2 cannot include a Placebo arm. In Scenario 2, comparison of A versus B in Trial 2 is of primary interest, sometimes followed by the comparison of Treatment B versus Placebo to indirectly infer efficacy of Treatment B through a cross-trial comparison.

In Scenario 2 a new treatment is compared with an established treatment with the objective of demonstrating that new treatment is non-inferior to this established treatment.

The methodologies for making indirect cross-trial comparisons are available, for example, Refs. [31–34, 34–43]. The validity of these methods relies on strong assumptions that often cannot be formally tested because treatments are not compared directly within the same trial [39, 40].

3.2.1. Type I error and setting the non-inferiority limit

Choice of Type I error

Two simultaneous one-tailed tests setting $\alpha = 0.025$ would maintain an overall Type I error rate of 2.5%. However, the choice of the Type I error is a controversial issue. The convention for equivalence or non-inferiority trials is to set the Type I error rate at half of that which would be employed for a two-sided test used in a superiority trial, that is, $\alpha = 0.025$ [5]. Setting the Type I error rate for equivalence or non-inferiority trials at half that for superiority trials could be considered to be consistent. This is because although in a superiority trial we use a two-sided 5% significance level, in practice for most trials what we have is a one-sided investigation with a 2.5% level of significance. The reason for this is that we usually have an investigative therapy and a control therapy and it is only statistical superiority of the investigative therapy that is of interest.

Through the rest of the sections on equivalence and non-inferiority trials we will assume that $\alpha = 0.025$ and that 95% confidence intervals will be used in the final statistical analysis.

Choice of non-inferiority limit

We have already discussed the setting of non-inferiority limits but in general the following points should be considered:

(1) You must be confident that the active control would have been different from the placebo had one been employed.
(2) You should be able to determine that there is no clinically meaningful difference between the investigative treatment and the control treatment.
(3) Through comparing the investigative treatment to the control treatment you should indirectly be able to determine that it is superior to placebo.

Steps 1 and 3 are important because there is a view that non-inferiority and equivalence (discussed later in the paper) trials reward ‘failed’ studies, that is, if we conducted a poor trial where it would not have been possible to demonstrate the control treatment to be superior to placebo, then a poor investigative therapy may be accepted by comparison to this control. However, Julious and Zariffa [44] pointed out that this may not be the case because poor studies are poor for most objectives as poor studies tend to have higher statistical variability and so are less likely therefore to show non-inferiority or equivalence.

We can therefore infer that the clinical difference used for the limits of equivalence and non-inferiority will be smaller than the difference used for placebo controlled superiority trials. There also is no generic definition for its setting — its definition will need to be defined on a study-by-study or indication-by-indication basis with consultation with the appropriate agencies and experts.

There are regulatory guidelines for a binary response in the antimicrobial therapeutic area where active controlled trials are the norm [45, 46]. The issues raised from this therapeutic area are generic to other therapeutic areas.

Table XII gives the non-inferiority margins for different response rates as recommended by the FDA [45] and CHMP [46]. The FDA guidelines are redundant now but they do raise interesting points. What is evident from Table XII is that although the CPMP recommends a flat non-inferiority margin, the FDA margins are a step function according to the anticipated control response rate.

3.2.2. Sample size calculation. The issue in calculating the sample size is that under both the null and alternative there is a non-zero difference between treatments. Generally, sample size formulas can be

Table XII. Non-inferiority margins for different control response rates.

	Non-inferiority margin	
Response rate	FDA*	CHMP†
$\geqslant 90$	−10%	−10%
80–89%	−15%	−10%
70–79%	−20%	−10%

*Food and Drug Authority.

†Committee for Health and Medicinal Products (CHMP; formerly Committee for Pharmaceutical and Medicinal Products (CPMP)).

thought of as Equation (27).

$$n_{\mathrm{A}} = \frac{\left(Z_{1-\alpha}\sqrt{\text{Variance under Null}} + Z_{1-\beta}\sqrt{\text{Variance under the Alternative}}\right)^2}{((\pi_{\mathrm{A}} - \pi_{\mathrm{B}}) - d)^2}. \tag{27}$$

Now (27) can be written as

$$n_{\mathrm{A}} = \frac{\left(Z_{1-\alpha}\sqrt{\tilde{\pi}_{\mathrm{A}}(1-\tilde{\pi}_{\mathrm{A}}) + \tilde{\pi}_{\mathrm{B}}(1-\tilde{\pi}_{\mathrm{B}})} + Z_{1-\beta}\sqrt{\pi_{\mathrm{A}}(1-\pi_{\mathrm{A}}) + \pi_{\mathrm{B}}(1-\pi_{\mathrm{B}})}\right)^2}{((\pi_{\mathrm{A}} - \pi_{\mathrm{B}}) - d)^2}, \tag{28}$$

where $\tilde{\pi}_{\mathrm{A}}$ and $\tilde{\pi}_{\mathrm{B}}$ are estimates of the responses on treatment under the null hypothesis used to estimate the variance under this hypothesis. For non-inferiority trials we have that $\pi_{\mathrm{A}} \neq \pi_{\mathrm{B}}$, that is, the two treatments do not have an equal response. Because the estimates of π_{A} and π_{B} affect the estimate of the variance, the definition of the null hypothesis hence influences the variance under this hypothesis. There are a number of ways of considering this problem, three of which will now be discussed [4, 47–50]. Julious and Owen [51] compared the different methods through simulation and within the parameters of the simulation recommended the simplest method for sample size estimation was to estimate the variance under the null hypothesis simply by replacing $\tilde{\pi}_{\mathrm{A}}$ and $\tilde{\pi}_{\mathrm{B}}$ with anticipated estimates of the response, π_{A} and π_{B}. Hence, the variance of a single observation under the null hypothesis becomes

$$\pi_{\mathrm{A}}(1-\pi_{\mathrm{A}}) + \pi_{\mathrm{B}}(1-\pi_{\mathrm{B}}), \tag{29}$$

which is the same as the variance under the alternative.

For the special case of equal sized groups, that is, $n_{\mathrm{A}} = n_{\mathrm{B}}$, a direct estimate of the sample size can be obtained [47]

$$n_{\mathrm{A}} = \frac{(\pi_{\mathrm{A}}(1-\pi_{\mathrm{A}}) + \pi_{\mathrm{B}}(1-\pi_{\mathrm{B}}))(Z_{1-\beta} + Z_{1-\alpha})^2}{((\pi_{\mathrm{A}} - \pi_{\mathrm{B}}) - d)^2}. \tag{30}$$

where π_{A} is the assumed proportion of responses expected in subjects on treatment A and π_{B} is the assumed proportion of responses in subjects on treatment B. Table XIII gives sample size estimates for 90% power and a Type I error rate of 2.5%.

As we discussed with superiority trials, Equation (30) could be adapted to give the maximum sample size for the cases where $\overline{\pi} = 0.5$ (where $\overline{\pi} = (\pi_{\mathrm{A}} + \pi_{\mathrm{B}})/2$) [4]. Hence, a quick estimate of the sample size, for 90% power and two-sided significance level of 5%, can be obtained from the following result:

$$n_{\mathrm{A}} = \frac{5.25}{((\pi_{\mathrm{A}} - \pi_{\mathrm{B}}) - d)^2}. \tag{31}$$

While for 80% power and two-sided significance level of 5% the sample size can be estimated from

$$n_{\mathrm{A}} = \frac{4}{((\pi_{\mathrm{A}} - \pi_{\mathrm{B}}) - d)^2}. \tag{32}$$

Both of these results will provide conservative ‘maximum’ estimates of the sample size. The utility of these results here could be questioned however, because often with non-inferiority trials the anticipated

Table XIII. Sample sizes for a non-inferiority study for 90% power and a Type I error rate of 2.5%.

		$\pi_B - \pi_A$										
π_A	Limit	−0.05	−0.04	−0.03	−0.02	−0.01	0	0.01	0.02	0.03	0.04	0.05
0.70	0.05	—	45845	11325	4993	2784	1766	1214	883	669	522	418
0.70	0.10	1839	1268	925	703	550	442	362	301	254	216	186
0.70	0.15	460	378	315	266	228	197	171	150	133	118	105
0.70	0.20	205	179	157	139	124	111	100	90	81	74	67
0.75	0.05	—	41537	10222	4491	2495	1577	1080	782	590	459	366
0.75	0.10	1671	1149	835	632	493	395	322	267	224	190	163
0.75	0.15	418	342	284	240	204	176	152	133	117	103	92
0.75	0.20	186	162	142	125	111	99	89	80	72	65	59
0.80	0.05	—	36178	8856	3872	2141	1345	917	660	495	382	303
0.80	0.10	1461	1000	723	545	423	337	273	225	188	158	135
0.80	0.15	366	298	246	207	175	150	129	112	98	86	76
0.80	0.20	163	141	123	108	95	85	75	67	60	54	49
0.85	0.05	—	29768	7227	3136	1720	1072	724	516	383	293	229
0.85	0.10	1209	822	590	441	340	268	216	176	145	121	102
0.85	0.15	303	245	201	167	141	120	102	88	76	66	58
0.85	0.20	135	116	101	88	77	67	60	53	47	42	37
0.90	0.05	—	22308	5336	2284	1234	757	502	351	255	190	145
0.90	0.10	915	615	436	322	244	190	150	120	97	79	65
0.90	0.15	229	183	149	122	101	85	71	60	51	43	37
0.90	0.20	102	87	74	64	55	48	41	36	31	27	24

Table XIV. Sample sizes for a non-inferiority study for 90% power and a Type I error rate of 2.5% for a control response rate (π_A) of 95%.

	$\pi_B - \pi_A$						
Limit	−0.03	−0.02	−0.01	0	0.01	0.02	0.03
0.03	—	11964	2780	1146	585	336	208
0.04	12904	3020	1249	655	386	242	156
0.05	3249	1358	717	424	271	184	129

responses are likely to be high on both treatment arms and results (31) and (32) are very conservative outside of the range (0.3, 0.7) for $\bar{\pi}$.

Sample size estimates using Equation (30) are given in Table XIII for the range $0.70 \leqslant \pi_A \leqslant 0.90$ to illustrate the issues with non-inferiority sample size calculations. Note that how for a trial being designed where the new treatment is thought to be a little better than control, that is, $\pi_B - \pi_A > 0$, the sample size is smaller than for $\pi_B - \pi_A = 0$. The opposite is true for $\pi_B - \pi_A < 0$.

Sample sizes are not given for anticipated responses greater than 0.90 as for high response rates the Normal approximation used in the sample size calculations may no longer hold. Our recommendation for sample sizes outside of this range would be to estimate the values using alternative methods such as simulation, which we describe below.

Table XIV gives an example of sample size calculations where the control response is assumed to be 95% for various non-inferiority limits and true mean differences. The process for the simulation was as follows:

(1) Simulate a random sample of size n from a binomial distribution, where the response rates in the two arms are assumed to be $\pi_A = 0.95$ and π_B goes from 0.92 to 0.98.
(2) For the random sample estimate the response rates in the two treatment arms.
(3) Calculate a 95% confidence interval for the treatment difference $\pi_A - \pi_B$ and determine if the lower bound excludes the non-inferiority limit.

(4) Repeat 1 to 3 a large number of times (here 100,000) and count the number of times simulations conclude non-inferiority. Take this as the power for the sample size.
(5) Repeat 1 to 4 increasing the sample size by 1 until a cutoff for the power has been reached.

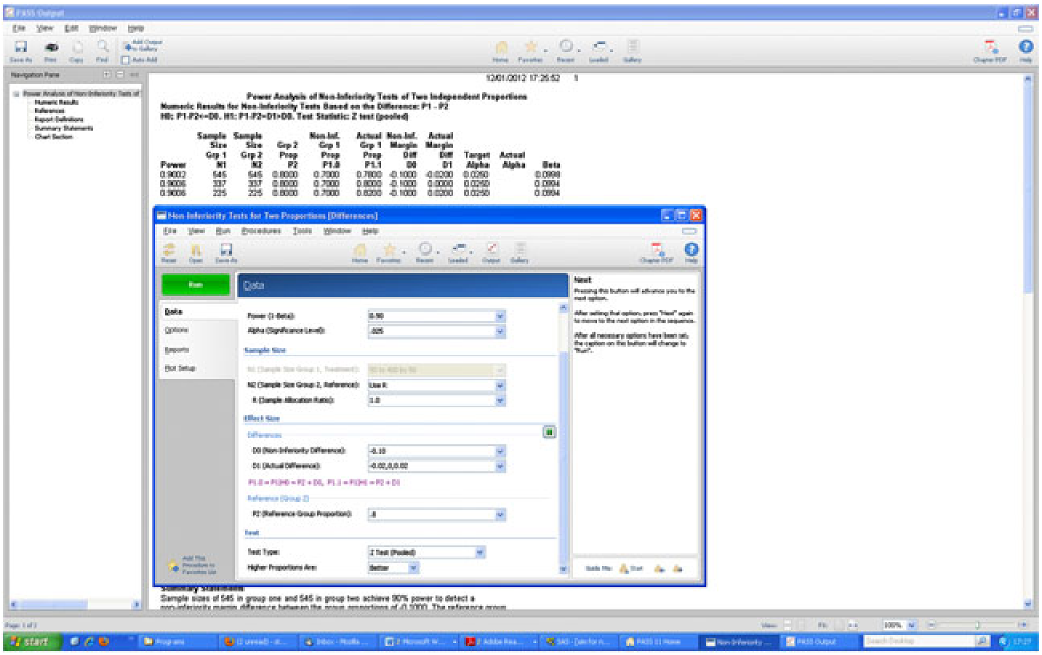
The sample sizes were simulated beyond 90% power so we can confirm that the study with the given sample size and all proceeding sample sizes (up to 10 greater) had 90% power. The confidence intervals for the simulation were calculated using the Wilson score method. This methodology has been shown to perform well in simulations and to give comparable results to exact methodologies [52]. We assume that the confidence intervals in the analysis in the completed study would be the Wilson score ones.

3.2.3. Worked Example 4 — sample size calculation for a parallel group non-inferiority trial with binary response. An investigator wishes to design a trial where the anticipated response rate on the active control is 80%. The investigator also expects an 80% response rate on the investigative therapy, that is, the investigator anticipates there to be no true difference between the treatments. The non-inferiority limit is to be set at 10%, the sample size is to be estimated with 90% power and a one-sided Type I error rate of 2.5%.

From Equation (30) the sample size is estimated to be 337 patients per arm. If the investigative response rate was anticipated to be 82% — a little better than the control response — then the sample size would be reduced to 225 patients per arm. Hence, only a small difference between treatments can have quite a marked effect on the sample size.

If the investigator thought that the investigative treatment is a little worse, say 78% rather than 80% then for the same non-inferiority limit the sample size is increased to 545 patients per arm. This demonstrates that a small change in the investigative response rate gives a substantial increase in the sample size.

To undertake the calculation in PASS, there are a number of options. One route is under menu to select ‘Proportions/Two Groups: Independent/Non-inferiority [Differences]’. Then in the dialogue screen: for ‘Test Statistic’ select ‘Z-test Pooled’; for ‘Non-inferiority Difference’ enter –0.10; for ‘Actual Difference’ enter –0.02, 0 and 0.02; and for ‘Reference Group Proportion’ enter 0.80. Example output from PASS is given below. PASS for this example gives the same sample size estimates as given in Table XIII.



```
                                                            12/01/2012 17:26:52    1

                Power Analysis of Non-Inferiority Tests of Two Independent Proportions
Numeric Results for Non-Inferiority Tests Based on the Difference: P1 - P2
H0: P1-P2<=D0. H1: P1-P2=D1>D0. Test Statistic: Z test (pooled)

         Sample  Sample          Non-Inf.  Actual  Non-Inf.   Actual
           Size    Size   Grp 2     Grp 1   Grp 1   Margin    Margin
          Grp 1   Grp 2    Prop      Prop    Prop     Diff      Diff   Target  Actual
Power        N1      N2      P2      P1.0    P1.1       D0        D1    Alpha   Alpha     Beta
0.9002      545     545  0.8000    0.7000  0.7800  -0.1000   -0.0200   0.0250           0.0998
0.9006      337     337  0.8000    0.7000  0.8000  -0.1000    0.0000   0.0250           0.0994
0.9006      225     225  0.8000    0.7000  0.8200  -0.1000    0.0200   0.0250           0.0994
```

Non-Inferiority Tests for Two Proportions [Differences]

Data

Power (1-Beta): 0.90
Alpha (Significance Level): .025

Sample Size

N1 (Sample Size Group 1, Treatment): 50 to 400 by 50
N2 (Sample Size Group 2, Reference): Use R
R (Sample Allocation Ratio): 1.0

Effect Size

Differences

D0 (Non-Inferiority Difference): -0.10
D1 (Actual Difference): -0.02,0,0.02

P1.0 = P1|H0 = P2 + D0, P1.1 = P1|H1 = P2 + D1

Reference (Group 2)

P2 (Reference Group Proportion): .8

Test

Test Type: Z Test (Pooled)
Higher Proportions Are: Better

Next

Pressing this button will advance you to the next option.

After setting that option, press "Next" again to move to the next option in the sequence.

After all necessary options have been set, the caption on this button will change to "Run".

Summary Statements

Sample sizes of 545 in group one and 545 in group two achieve 90% power to detect a non-inferiority margin difference between the group proportions of -0.1000. The reference group

InnQuery for ‘Making Conclusions Using’ tick ‘Proportions’; for ‘number of Groups’ tick ‘Two’; for ‘Analysis Method’ tick ‘Equivalence’ then select ‘Two group test of equivalence in proportions’. nQuery also agrees with both PASS and the sample size estimates from Table XIII for this worked example.

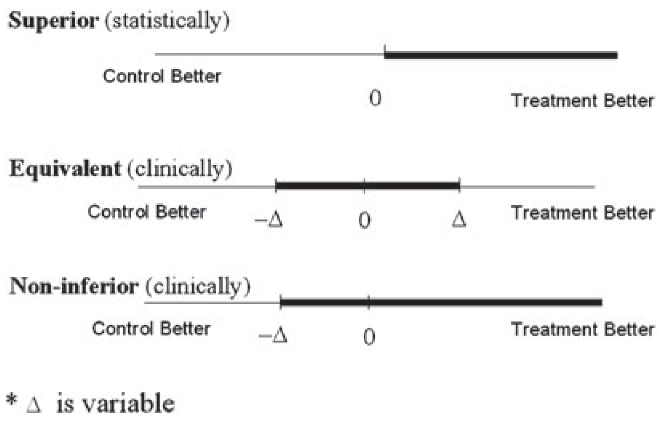


Figure 6. An illustration of the difference between superiority, equivalence and non-inferiority trials: the dark line in the figure is the confidence interval while delta is the non-inferiority or equivalence limit

3.3. *'As good as or better' trials*

For certain clinical trials the objective is to demonstrate either that a given treatment is clinically not inferior or that it is clinically superior when compared with the control, that is, that the treatment is 'as good as or better' than the control. In 'as good as or better' trials two null hypotheses are investigated. First, the non-inferiority null hypothesis:

H_0: A given treatment is inferior with respect to the mean response.
If this null hypothesis is rejected then a second null hypothesis can be investigated:
H_0: The two treatments have equal effect with respect to the mean response.

In practice these two null hypotheses are investigated through the construction of a 95% confidence interval to investigate where the lower (or upper as appropriate) bound lies. Figure 6 highlights how the two separate hypotheses for superiority and non-inferiority are investigated.

It should be noted that 'as good as or better' trials are really a subcategory of either superiority or non-inferiority trials. However, we have to put them into a separate section to highlight how they combine the null hypotheses of superiority and non-inferiority trials into one closed testing procedure while maintaining the overall Type I error.

To introduce the closed testing procedure, we will first describe the situation where a one-sided test of non-inferiority is followed by a one-sided test of superiority. The more general case where a one-sided test of non-inferiority is followed by a two-sided test of superiority is then described.

This section is given in Julious [1] where the calculations for data anticipated to have a Normal distribution are given. This drew on the work of Morikawa and Yoshida [53]. The CPMP have a 'points to consider' document on the topic [28].

3.3.1. *A test of non-inferiority and a one sided test of superiority.*

The null ($H1_0$) and alternative ($H1_1$) hypotheses for a non-inferiority trial can be written as:

$H1_0$: $\pi_A - \pi_B \leqslant -d$.
$H1_1$: $\pi_A - \pi_B > -d$.

This can alternatively be written as

$H1_0$: $\pi_A - \pi_B + d \leqslant 0$.
$H1_1$: $\pi_A - \pi_B + d > 0$.

The corresponding null ($H2_0$) and alternative ($H2_1$) hypotheses for a superiority trial can be written as:

$H2_0$: $\pi_A - \pi_B \leqslant 0$.
$H2_1$: $\pi_A - \pi_B > 0$.

What is clear from the definitions of these hypotheses is that if $H2_0$ is rejected at the α level then $H1_0$ would also be rejected. Also, if $H1_0$ is not rejected at the α level then $H2_0$ would also not be rejected.

This is because $\pi_A - \pi_B + d \geqslant \pi_A - \pi_B$. Hence, both $H1_0$ and $H2_0$ are rejected if they are both statistically significant; neither $H1_0$ and $H2_0$ are rejected if $H1_0$ is not significant; and only $H1_0$ is rejected if only $H1_0$ is significant.

On the basis of these properties a closed test procedure can be applied to investigate both non-inferiority and superiority while maintaining the overall Type I error rate without α adjustment. To do this, the intersection hypothesis $H2_0 \cap H1_0$ is first investigated which, if rejected, is followed by a test of $H1_0$ and $H2_0$. In this instance $H2_0 \cap H1_0 = H1_0$ and so both non-inferiority and superiority can be investigated through the following two steps [1].

(1) First investigate the non-inferiority through the hypothesis $H1_0$. If $H1_0$ is rejected then $H2_0$ can be tested. If $H1_0$ is not rejected then the investigative treatment is inferior to the control treatment.
(2) If $H2_0$ is then rejected in the next step one can conclude that the investigative treatment is superior to the control. Otherwise, if $H20_o$ is not rejected then non-inferiority should be concluded.

3.3.2. A test of non-inferiority and a two sided test of superiority. The null ($H3_0$) and alternative ($H3_1$) hypotheses for a two-sided test of superiority can be written as:

$H3_0$: $\pi_A = \pi_B$.
$H3_1$: $\pi_A < \pi_B$ or $\pi_A > \pi_B$.

The hypothesis $H3_1$ is equivalent to two one-sided tests at the $\alpha/2$ level of significance (summing to give an overall Type I error rate of α) through the investigation of $H2_0$ against the alternative of $H2_1$ and the following null and alternative hypotheses:

$H4_0$: $\pi_A \geqslant \pi_B$.
$H4_1$: $\pi_A < \pi_B$.

It is apparent that the intersection hypothesis $H1_0 \cap H3_0$ is always rejected because it is empty and so both $H1_0$ and $H3_0$ can be tested. Because of there being no intersection the following steps can be applied:

(1) If the observed treatment difference is greater than zero and $H3_0$ is rejected, then $H1_0$ is also rejected and one can conclude that the investigative treatment is statistically superior to control.
(2) If the observed treatment difference is less than zero and $H3_0$ is rejected and $H1_0$ is not, then the control is statistically superior to the investigative treatment. If $H1_0$ is also rejected then the investigative drug is worse than the control but is not inferior (practically although this may be difficult to claim).
(3) If $H3_0$ is not rejected but $H1_0$ is, then the investigative treatment is non-inferior compared with the control.
(4) If neither $H1_0$ nor $H3_0$ are rejected then one must conclude that the investigative treatment is inferior to control.

Note that when investigating the $H1_0$ and $H3_0$ hypotheses, $H3_0$ will be tested at a two-sided α level of significance while $H1_0$ will be tested at a one-sided $\alpha/2$ level of significance. Thus, the overall level of significance is maintained at α.

3.3.3. Non-inferiority versus superiority trials. Because non-inferiority trials often use a non-inferiority margin, which is set at a fraction of the superiority effect of the active control over placebo, the sample size requirements for a non-inferiority trial are often perceived as being much greater than that for a superiority trial. However, the sample size formulas are the same only when the non-inferiority margin is set to zero.

If the margin is set to zero it would mean that when we compare two active treatments the objective would be to show that the lower bound of the 95% confidence interval excludes zero, and the investigative treatment is statistically superior to the active control. A non-inferiority margin is usually set at less than zero. In this case it is therefore easier to show a new treatment is non-inferior and, in the active control trial context, this requires smaller sample sizes.

There is a further important distinction between superiority trials and non-inferiority trials in that the former use the data 'as randomised' and the principle of 'intention-to-treat'. For a non-inferiority trial it has been suggested that one should analyse the data 'per protocol' and also 'as randomised' as co-primary [28]. This may require that a greater number of subjects are recruited.

Table XV. Sample sizes per group for a superiority study for 90% power and various Type I error rates.

		Significance level							
π_A	$\pi_B - \pi_A$	0.02	0.05	0.07	0.10	0.12	0.15	0.17	0.20
0.70	0.02	10820	8818	7624	6766	6090	5534	5058	4642
0.70	0.03	4758	3878	3354	2976	2678	2434	2224	2042
0.70	0.04	2648	2158	1866	1656	1490	1354	1238	1136
0.70	0.05	1676	1366	1180	1048	944	856	784	718
0.75	0.02	9584	7812	6754	5992	5396	4902	4480	4112
0.75	0.03	4198	3422	2958	2624	2364	2146	1962	1802
0.75	0.04	2326	1896	1638	1454	1310	1190	1088	998
0.75	0.05	1466	1194	1032	916	824	750	684	628
0.80	0.02	8086	6592	5698	5056	4552	4136	3780	3470
0.80	0.03	3520	2870	2480	2202	1982	1800	1646	1510
0.80	0.04	1938	1580	1366	1212	1092	992	906	832
0.80	0.05	1212	988	854	758	682	620	568	520
0.85	0.02	6326	5156	4458	3956	3562	3236	2958	2714
0.85	0.03	2726	2222	1922	1704	1534	1394	1274	1170
0.85	0.04	1484	1210	1046	928	836	760	694	638
0.85	0.05	918	748	648	574	518	470	430	394
0.90	0.02	4302	3508	3032	2690	2422	2200	2012	1846
0.90	0.03	1816	1480	1280	1136	1022	928	848	780
0.90	0.04	966	788	680	604	544	494	452	414
0.90	0.05	582	474	410	364	328	298	272	250

The concepts of superiority and non-inferiority are of course interrelated. Indeed there may be instances where instead of designing a study to show an investigative treatment is no worse than an active control at the 2.5% level of significance we may wish to design a superiority study but at a level of statistical significance greater than the nominal two-sided 5% (one-sided 2.5%). Such a study would give more assurance as to the investigative treatment being no worse than the active control. CHMP [34] comment:

> *'It might be an acceptable approach, in extreme situations, to run a superiority trial using a less stringent significance level than P = 0.05, weighing up the increased risk of a false positive result against the risk of rejecting a drug with a valuable efficacy advantage. It might be more acceptable, and easier from an ethical perspective, to specify a level of confidence we require in the superiority of a drug, than to specify an extra number of deaths that is of no clinical importance...*
>
> *For example with a data-set where the lower bound of an 85% confidence interval (by definition narrower than a 95% interval) touches zero, it might be that the 95% interval touches –5. If delta had been defined to be –5 then achieving non-inferiority in this example would correspond to having demonstrated superiority at the 15% level of significance'*

Table XV gives sample sizes for different control response rates and different improvements on the investigative treatment, assuming $\pi_B > \pi_A$, for various control response rates and Type I error rates using result (17). Results (22) and (24) — exact methodology — could also be used for these calculations.

3.4. Equivalence trials

In certain cases the objective of a clinical trial is not to demonstrate that an investigative treatment is superior or no worse than a control but instead to demonstrate that two treatments have no clinically meaningful difference, that is, that they are clinically equivalent. The null (H_0) and alternative (H_1) hypotheses for such equivalence trials take the form:

H_0: The two treatments are different with respect to the risk difference ($\pi_A \neq \pi_B$).
H_1: The two treatments are not different with respect to the risk difference ($\pi_A = \pi_B$).

Usually these hypotheses are written in terms of a clinical difference, d, and become:

H_0: $\pi_A - \pi_B \leqslant -d$ or $\pi_A - \pi_B \geqslant +d$.
H_1: $-d < \pi_A - \pi_B < +d$.

These hypotheses are an example of an intersection–union test, in which the null hypothesis is expressed as a union and the alternative as an intersection. To conclude equivalence, one needs to reject each component of the null hypothesis. Note that in an intersection–union test, each component is tested at level α giving a composite test, which is also of level α [1,4,54].

A common approach with equivalence trials to test each component of the null hypothesis with a t-test: called the Two One-Sided Test procedure. In practice, this is operationally the same as constructing a $(1-2\alpha)100\%$ confidence interval where equivalence is concluded provided that each end of the confidence interval falls completely within the interval $(-d, +d)$ [30]. This is because the $(1–2\alpha)100\%$ confidence interval is excluding two regions each of size α. Hence, the overall significance level is α.

3.4.1. Sample sizes for an equivalence trial. The power for a given sample size can be estimated from

$$1-\beta = \Phi\left(\sqrt{\frac{n_A\left((\pi_A-\pi_B)-d\right)^2}{\pi_A(1-\pi_A)+\pi_B(1-\pi_B)}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{n_A\left((\pi_A-\pi_B)+d\right)^2}{\pi_A(1-\pi_A)+\pi_B(1-\pi_B)}} - Z_{1-\alpha}\right) - 1. \qquad (33)$$

To estimate the sample size one iterates (33) on the sample size until the nominal power is reached. Similar to non-inferiority trials discussed earlier, Equation (33) uses the anticipated responses in the trial to estimate the sample size and similar issues occur with respect to estimating the response on the null and alternative hypothesis [4,47,48]. Table XVI gives the sample sizes for various control response rates and equivalence limits.

Result (33) can be simplified for the case where there is a non-zero difference between treatments such that $\pi_A > \pi_B$. In this instance most of the Type II error comes from just one part of (33) and so a direct estimate of the sample size can be estimated by rewriting (33) as

Table XVI. Sample sizes per group for an equivalence study estimated for 90% power and a type I error rate of 2.5%.

		$\pi_B - \pi_A$										
π_A	Limit	−0.05	−0.04	−0.03	−0.02	−0.01	0	0.01	0.02	0.03	0.04	0.05
0.70	0.05	—	45645	11325	4993	2802	2184	2749	4806	100694	42282	—
0.70	0.10	1839	1268	925	707	585	546	574	680	874	1175	1671
0.70	0.15	460	378	317	275	252	243	247	265	299	350	418
0.70	0.20	205	180	161	148	140	137	138	143	152	167	186
0.75	0.05	—	41337	10222	4491	2511	1950	2445	4257	9434	37134	—
0.75	0.10	1671	1149	835	636	525	488	511	603	771	1032	1461
0.75	0.15	418	342	286	248	226	217	220	235	264	308	366
0.75	0.20	186	163	145	133	126	122	122	126	134	146	163
0.80	0.05	—	35978	8856	3872	2154	1664	2075	3592	7910	30934	—
0.80	0.10	1461	1000	723	548	450	416	434	509	646	860	1209
0.80	0.15	366	298	248	214	194	185	187	198	222	256	303
0.80	0.20	163	142	126	115	108	104	104	107	113	122	135
0.85	0.05	—	29568	7227	3136	1731	1326	1639	2809	6124	2368	—
0.85	0.10	1209	822	590	444	362	332	343	398	500	658	915
0.85	0.15	303	245	202	173	156	148	148	155	172	196	229
0.85	0.20	135	117	103	93	87	83	82	84	87	94	102
0.90	0.05	—	22108	5336	2284	1242	93	1136	1911	4075	1538	—
0.90	0.10	915	615	436	324	260	234	238	271	333	428	578
0.90	0.15	229	183	150	126	112	104	102	106	114	128	145
0.90	0.20	102	87	76	68	62	59	57	57	58	61	65

Table XVII. Sample sizes per group for an equivalence study estimated for 90% power and a Type I error rate of 2.5% estimated directly for a nonzero difference between treatments.

		$\pi_B - \pi_A$										
π_A	Limit	−0.05	−0.04	−0.03	−0.02	−0.01		0.01	0.02	0.03	0.04	0.05
0.70	0.05	—	45645	11325	4993	2784	1766	2732	4806	10694	42282	—
0.70	0.10	1839	1268	925	703	550	442	540	676	873	1175	1671
0.70	0.15	460	378	315	266	228	197	223	256	298	350	418
0.70	0.20	205	179	157	139	124	111	122	134	149	166	186
0.75	0.05	—	41337	10222	4491	2495	1577	2430	4257	9434	37134	—
0.75	0.10	1671	1149	835	632	493	395	480	599	771	1032	1461
0.75	0.15	418	342	284	240	204	176	199	227	263	307	366
0.75	0.20	186	162	142	125	111	99	108	119	131	146	163
0.80	0.05	—	35978	8856	3872	2141	1345	2062	3592	7910	30934	—
0.80	0.10	1461	1000	723	545	423	337	408	506	646	860	1209
0.80	0.15	366	298	246	207	175	150	169	192	220	256	303
0.80	0.20	163	141	123	108	95	85	92	100	110	121	135
0.85	0.05	—	29568	7227	3136	1720	1072	1628	2809	6124	23684	—
0.85	0.10	1209	822	590	441	340	268	322	396	500	658	915
0.85	0.15	303	245	201	167	141	120	133	150	171	196	229
0.85	0.20	135	116	101	88	77	67	73	79	85	93	102
0.90	0.05	—	22108	5336	2284	1234	757	1129	1911	4075	15383	—
0.90	0.10	915	615	436	322	244	190	223	269	333	428	578
0.90	0.15	229	183	149	122	101	85	93	102	114	128	145
0.90	0.20	102	87	74	64	55	48	51	54	57	61	65

$$n_A = \frac{\left(\pi_A(1-\pi_A) + \pi_B(1-\pi_B)\right)\left(Z_{1-\beta} + Z_{1-\alpha}\right)^2}{\left(|\pi_A - \pi_B| - d\right)^2}. \tag{34}$$

The greater π_A is relative to π_B the nearer $\pi_B - \pi_A$ is to the margin and the closer (34) becomes to (33). To illustrate this point, Table XVII estimates the sample sizes from (34). We can see here that the sample sizes approach those of Table XVI as the difference between the treatments gets bigger.

For the special case of no anticipated treatment difference the power can be estimated from

$$1 - \beta = 2\Phi\left(\sqrt{\frac{n_A d^2}{2\overline{\pi}(1-\overline{\pi})}} - Z_{1-\alpha}\right) - 1, \tag{35}$$

where $\overline{\pi} = (\pi_A + \pi_B)/2$ is interpreted in this instance as the anticipated overall response. Consequently (35) can in turn be rewritten to give a direct estimate of the sample size

$$n_A = \frac{2\left(Z_{1-\beta/2} + Z_{1-\alpha}\right)^2 \overline{\pi}(1-\overline{\pi})}{d^2}. \tag{36}$$

Hence, for the special case of no treatment difference, direct estimates of the sample size can be obtained.

3.4.2. Worked Example 5 — sample size calculation for a parallel group equivalence trial with binary response. An investigator wishes to design a trial where the anticipated response rate on the active control is 80%. The investigator also expects an 82% response rate on the investigative therapy, that is, there is anticipated to be a small difference between the treatments. The equivalence limit is to be set at 10%, the sample size is to be estimated with 90% power and a one-sided Type I error rate of 2.5%.

From Table XVI, the sample size is estimated to be 509 patients per arm. If the investigative response rate was also anticipated to be 80%, the sample size would be 416 patients per arm. If the response rate was expected to be 78% then the sample size would be estimated to be 548 patients per arm.

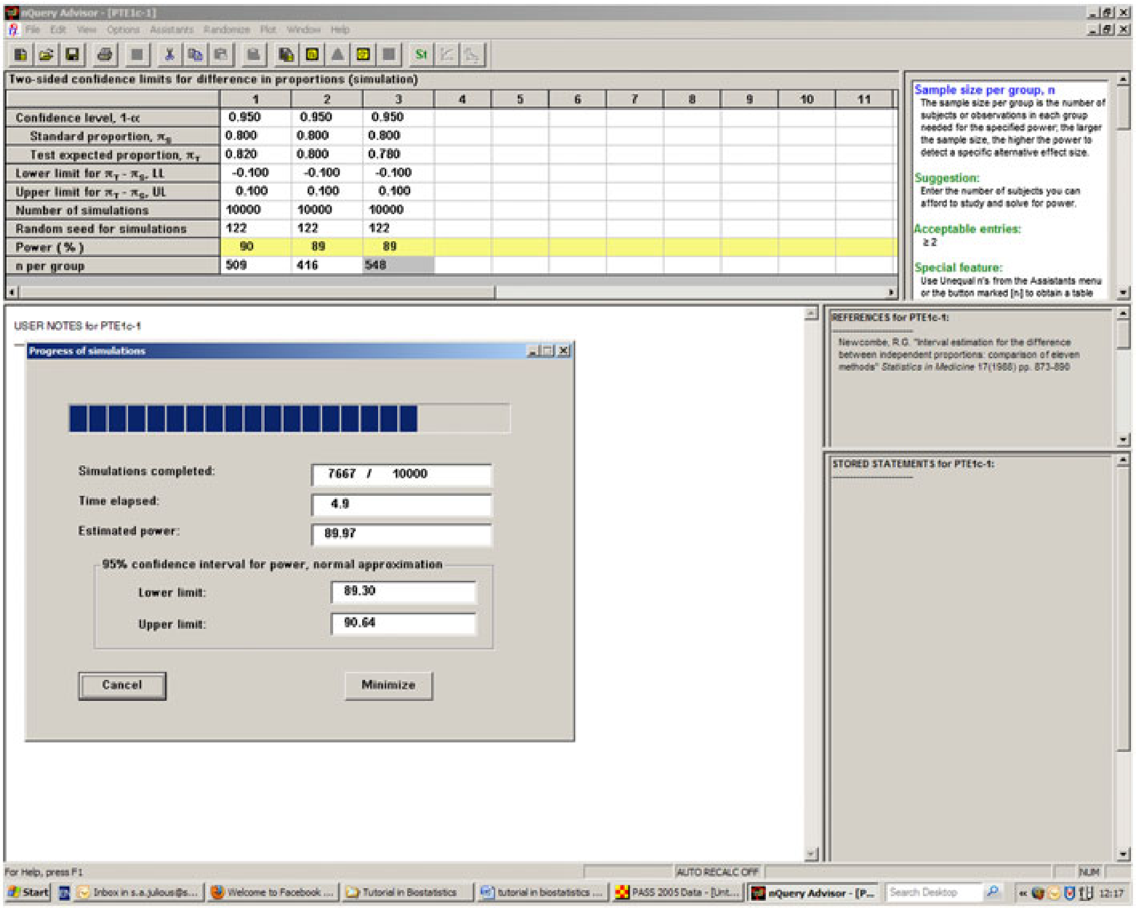
Equivalence trials are quite sensitive to the assumptions around the difference in responses especially as any non-zero difference will increase the sample size.

If we had used result (34) (and Table XVII) then we would have estimated the sample size to be 506 patients per arm. A little smaller than our previous sample size estimate.

To get the sample sizes per arm in nQuery for 'Making Conclusions Using' tick 'Proportions'; for 'number of Groups' tick 'Two'; for 'Analysis Method' tick 'Equivalence' then select 'Two group test of equivalence in proportions (using confidence interval)' and then 'Two sided confidence interval for test minus standard'. nQuery does not estimate the sample size directly but instead undertakes simulations to estimate the power for a given sample size. You then need to iterate to get the requisite power.

For the worked example described earlier of investigative and control responses of 82% and 80%, respectively, with 509 patients per arm taken from Table XVI, nQuery estimates the power to be 90%. For response rates of 80% on both arms nQuery estimates the power to be 89% with 416 patients per arm (417 patients per arm gives 90%). Finally, for response rates of 78% and 80%, nQuery estimates the power to be 89% with 548 patients per arm (for 550 patients nQuery estimates the power to be 90%).

An example nQuery output is given below.


nQuery Advisor - [PTE1c-1]
File Edit View Options Assistants Randomize Plot Window Help
St

Two-sided confidence limits for difference in proportions (simulation)

	1	2	3	4	5	6	7	8	9	10	11
Confidence level, 1-α	0.950	0.950	0.950								
Standard proportion, π_S	0.800	0.800	0.800								
Test expected proportion, π_T	0.820	0.800	0.780								
Lower limit for $\pi_T - \pi_S$, LL	-0.100	-0.100	-0.100								
Upper limit for $\pi_T - \pi_S$, UL	0.100	0.100	0.100								
Number of simulations	10000	10000	10000								
Random seed for simulations	122	122	122								
Power (%)	90	89	89								
n per group	509	416	548								

Sample size per group, n
The sample size per group is the number of subjects or observations in each group needed for the specified power; the larger the sample size, the higher the power to detect a specific alternative effect size.

Suggestion:
Enter the number of subjects you can afford to study and solve for power.

Acceptable entries:
≥ 2

Special feature:
Use Unequal n's from the Assistants menu or the button marked [n] to obtain a table

USER NOTES for PTE1c-1

Progress of simulations

Simulations completed: 7667 / 10000
Time elapsed: 4.9
Estimated power: 89.97

95% confidence interval for power, normal approximation
Lower limit: 89.30
Upper limit: 90.64

REFERENCES for PTE1c-1:

Newcombe, R.G. "Interval estimation for the difference between independent proportions: comparison of eleven methods" *Statistics in Medicine* 17(1998) pp. 873-890

STORED STATEMENTS for PTE1c-1:

For Help, press F1

There are a number of options to undertake the calculation in PASS. One route is under menu to select Proportions/Two Independent Proportions/Equivalence/Equivalence Test for Two Proportions [Differences]. Then in the dialogue screen: for 'Test Statistic' select 'Z-test Pooled'; for 'Upper Equivalence Difference' enter 0.10; for 'Actual Difference' enter –0.02, 0 and 0.02; and for 'Reference Group Proportion' enter 0.80. PASS agrees in the main with the sample size estimates taken from Table XVI except when the investigative and control responses were 78% and 80%, respectively, where it estimates the sample size to be 549 patients instead of 548 patients.

3.5. Estimation to a given precision

So far we have discussed specific defined objectives. However, there are cases when a preliminary, or pilot, investigation is conducted to estimate possible effects with a view to doing a later definitive study [55–57]. By definition, such studies are held early in the drug development (or clinical investigation) paradigm. With estimation studies, rather than formally testing a null hypothesis it is more informative to give a confidence interval for the unknown effect.

Precision calculations may also be undertaken when the sample size is determined primarily by practical considerations. In such cases one may quote the precision of the estimates obtained based on the half-width of the confidence interval, and provide this information in the discussion of the fixed sample size. Again it must be clearly stated in the protocol that the size of the study was determined based on practical, and not formal, considerations.

In the context of an overall clinical development (or investigation) an estimation study (or studies) could provide important cumulative evidence of the pharmacological benefit of a given drug asset. These studies cannot prove a given effect but can inform studies that can.

A conservative approach would be to set $\overline{p} = 0.5$ because if we do not have any idea of the overall response this would give us a maximum estimate of the variance for the absolute risk difference and would not be too conservative provided that $\overline{p}$ is within the range (0.3, 0.7). Therefore, for a given half confidence interval width, w, the following condition must be met to obtain the sample size per group

$$n_A = \frac{2\overline{p}(1-\overline{p})Z^2_{1-\alpha/2}}{w^2}. \tag{37}$$

where $\overline{p} = (pa + pb)/2$. Table XVIII is derived from Equation (37). Table XVIII gives the sample size required for different values of the expected mean response across treatment groups, $\overline{p}$, and widths, w. Two-sided 95% confidence intervals are planned in the final analysis. The mean responses, $\overline{p}$, given in the table vary from 0.10 to 0.50. Values greater than 0.50 are not given becauses the sample size required for $\overline{p} = 0.60$ is equivalent to $\overline{p} = 0.40$, the sample size for $\overline{p} = 0.70$ is the same as $\overline{p} = 0.30$, etc.

3.5.1. Worked Example 6 — sample size calculation for a parallel group estimation trial with binary response. An investigator wishes to design a trial where the average response rate is anticipated to be 65%. The investigator wishes to estimate possible effects with precision of +/–10% using a 95% confidence interval.

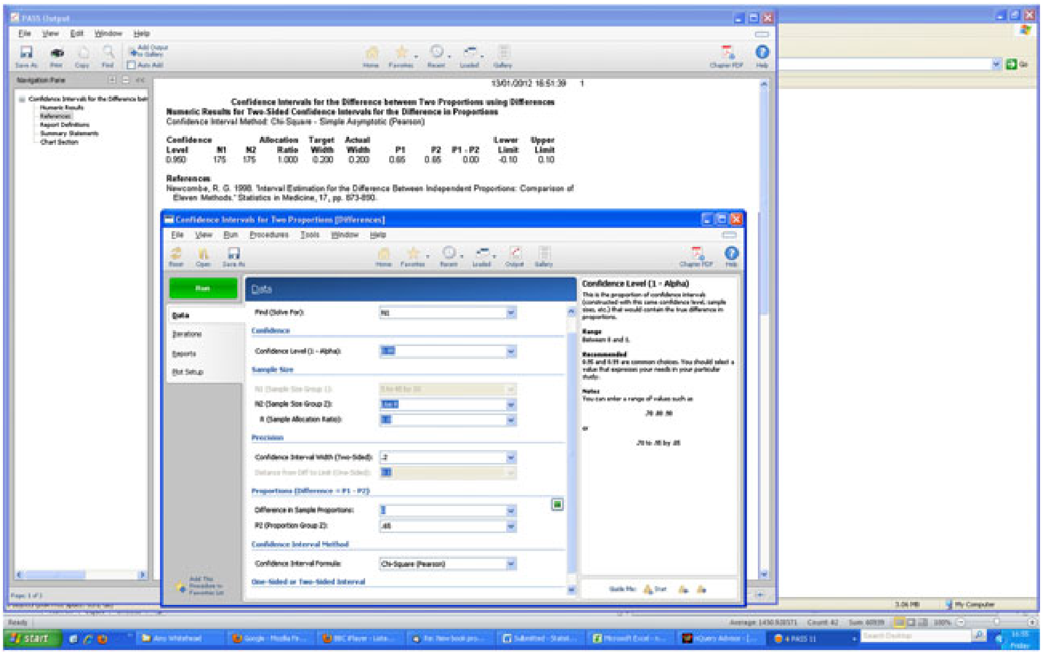
In Table XVIII we use $1 - \overline{p} = 0.35$ and get the sample size to be 175 patients per arm.

To estimate the sample sizes in nQuery for 'Making Conclusions Using' tick 'Proportions'; for 'number of Groups' tick 'Two'; for 'Analysis Method' tick 'Confidence Interval' then select 'Confidence for difference in proportions'. For an average response rate of 65% and precision of 10%, nQuery agrees with Table XVIII and estimates the sample size to be 175 patients per arm.

Table XVIII. Sample sizes required per group for two-sided 95% confidence intervals for different values of width, w, for various expected mean absolute responses.

	W				
$\overline{p}$	0.025	0.050	0.075	0.100	0.150
0.10	1107	277	123	70	31
0.15	1568	392	175	98	44
0.20	1967	492	219	123	55
0.25	2305	577	257	145	65
0.30	2582	646	287	162	72
0.35	2797	700	311	175	78
0.40	2951	738	328	185	82
0.45	3043	761	339	191	85
0.50	3074	769	342	193	86

In PASS, to calculate the sample size, you need to select ‘Confidence Intervals’ and the ‘Proportions’ and finally ‘Confidence Intervals for Two Proportions [Differences]’. For Confidence Interval Width set the value at 20% (note in this paper we use half widths); for P2 enter 0.65 and for ‘Confidence Interval’ ‘Chi-square’. PASS gives a sample size of 175 patients per arm.



```
                                                          13/01/2012 16:51:39    1

        Confidence Intervals for the Difference between Two Proportions using Differences
Numeric Results for Two-Sided Confidence Intervals for the Difference in Proportions
Confidence Interval Method: Chi-Square - Simple Asymptotic (Pearson)

Confidence              Allocation  Target  Actual                        Lower   Upper
Level        N1    N2       Ratio   Width   Width    P1     P2   P1 - P2  Limit   Limit
0.950       175   175       1.000   0.200   0.200   0.65   0.65     0.00  -0.10    0.10

References
Newcombe, R. G. 1998. 'Interval Estimation for the Difference Between Independent Proportions: Comparison of
  Eleven Methods.' Statistics in Medicine, 17, pp. 873-890.
```

Confidence Intervals for Two Proportions [Differences]

Data

Find (Solve For): N1

Confidence

Confidence Level (1 - Alpha): [illegible]

Sample Size

N2 (Sample Size Group 2): [illegible]

R (Sample Allocation Ratio): [illegible]

Precision

Confidence Interval Width (Two-Sided): .2

Proportions (Difference = P1 - P2)

Difference in Sample Proportions: [illegible]

P2 (Proportion Group 2): .65

Confidence Interval Method

Confidence Interval Formula: Chi-Square (Pearson)

One-Sided or Two-Sided Interval

Confidence Level (1 - Alpha)

This is the proportion of confidence intervals (constructed with this same confidence level, sample sizes, etc.) that would contain the true difference in proportions.

Range
Between 0 and 1.

Recommended
0.95 and 0.99 are common choices. You should select a value that expresses your needs in your particular study.

Notes
You can enter a range of values such as

.70 .80 .90

or

.70 to .90 by .05

4. Discussion

This paper describes sample size calculations when the outcome is binary for a variety of study designs. It is important to realise that sample size calculations are ‘a guess masquerading as mathematics’. Thus, we usually only need an approximate answer, and it is important that some form of sensitivity analysis is carried out to investigate what factors are important, and perhaps where more information should be sought.

References

1. Julious SA. Tutorial in Biostatistics: sample sizes for clinical trials with normal data. *Statistics in Medicine* 2004; **23**:1921–1986.
2. Sahai H, Khurshid A. Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two sample designs: a review. *Statistics in Medicine* 1996; **15**:1–21.
3. Chow SC, Shao J, Wang H. *Sample Size Calculations in Clinical Research*, 2nd ed. Chapman and Hall/CRC: Boca Raton, Fl, 2007.
4. Julious SA. *Sample Sizes for Clinical Trials*. Chapman and Hall: London, 2009.
5. ICH E9. Statistical principals for clinical trials, September 1998. Available at URL: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf (date last accessed: 1 Feb 2012).
6. ICH E3. Structure and content of clinical study reports, July 1996. Available at URL: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E3/Step4/E3_Guideline.pdf (date last accessed: 1 Feb 2012).
7. Hintze J. *PASS 11*. NCSS, LLC: Kaysville, UTAH, USA, 2011. www.ncss.com.
8. Elashoff JD. *nQuery Advisor Version 7 User’S Guide*. Statistical Solutions: Los Angeles, 2007.
9. Machin D, Campbell MJ, Tan SB, Tan SH. *Sample Size Tables for Clinical Studies*, 3rd ed. Wiley-Blackwell: Chichester, 2008.
10. Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not occurred: a statistical reminder. *BMJ* 1995; **311**:619–620.
11. Newcombe RG. Two sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.
12. Clopper CJ, Pearson ES. The use of confidence or fiduciallimits illustrated in the case of the binomial. *Biometrika* 1934; **26**:404–413.
13. Daly L. Simple SAS Macros for the calculation of exact binomial and Poisson confidence limits. *Computational and Biological Medicine* 1992; **22**:351–361.

14. Julious SA. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 2005; **24**:3383–3384.
15. SAS Institute Inc. *SAS 9.1.3 Help and Documentation*. SAS Institute Inc.: Cary, NC, 2002-2004.
16. Bland JM, Altman DG. Statistical notes: one sided and two sided tests of significance. *BMJ* 1994; **309**:248.
17. Fleming TR. One-sample multiple testing procedure for Phase II clinical trial. *Biometrics* 1982; **38**:143–151.
18. Korn EL. Sample size tables for bounding small proportions. *Biometrics* 1986; **42**(1):213–216.
19. Desu MM, Raghavarao D. *Sample Size Methodology*. Academic Press: London, 1990.
20. Julious SA, Campbell MJ, Altman DG. Estimating sample sizes for continuous, binary and ordinal outcomes in paired comparisons: practical hints. *Journal of Biopharmaceutical Statistics* 1999; **9**(2):241–251.
21. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal* 1995; **311**:1145–1148.
22. Ambrosius WT, Mahnken JD. Power for studies with random group sizes. *Statistics in Medicine* 2010; **29**:1137–1144.
23. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007; **6**(3):161–170.
24. Yeo A, Qu Y. Evaluation of the statistical power for multiple tests: a case study. *Pharmaceutical Statistics* 2009; **8**(1): 5–11.
25. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine* 2010. DOI: 10.1002/sim.3972.
26. Richardson JTE. The analysis of 2x1 and 2x2 contingency tables: an historical review. *Statistical Methods in Medical Research* 1994; **3**:107–133.
27. Lydersen S, Fagerland MW, Laake P. Tutorial in Biostatistics: recommended tests for association in 2×2 tables. *Statistics in Medicine* 2009; **28**:1159–1175.
28. CPMP. Points to consider on switching between superiority and non-inferiority. (CPMP/EWP/482/99), 17 July 2000. Available at URL: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf (date last accessed: 1 Feb 2012).
29. ICH E10 Choice of control group in clinical trials, 2000, May 2001. Available at URL: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002925.pdf (date last accessed: 1 Feb 2012).
30. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* 1996; **313**:36–39.
31. D'Agostino RB, Massaro J, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**:169–186.
32. Hung HMJ, Wang SJ, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:213–225.
33. Wiens BL. Choosing an equivalence limit for non-inferiority and or equivalence studies. *Controlled Clinical Trials* 2002; **23**:2–14.
34. CHMP Guideline on the choice of non-inferiority margin. Doc CPMP/EWP/2158/99, 2005. Available at URL: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf (date last accessed: 1 Feb 2012).
35. Wang SJ, Hung HMJ, Tsong Y. *Non-inferiority Analysis in Active Controlled Trials*, 2nd Edition, Encyclopaedia of Biopharmaceutical Statistics. Marcel Dekker: New York, NY; 674–677, 2003.
36. FDA Guidance for Industry, Non-Inferiority Clinical Trials (draft), March 2010. Available at URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf (date last accessed: 1 Feb 2012).
37. Julious SA. The ABC of non-inferiority margin setting from indirect comparisons. *Pharmaceutical Statistics* 2011. DOI: 10.1002/pst.517.
38. Julious SA, Wang SJ. Issues with indirect comparisons in clinical trials particularly with respect to non-inferiority trials. *Drug Information Journal* 2008; **42**(6):625–633.
39. Wang SJ, Hung HMJ, Tsong Y. Utility and pitfalls of some statistical methods in active controlled trials. *Controlled Clinical Trials* 2002; **23**:15–28.
40. Snapinn SM. Alternatives for discounting in the analysis of non-inferiority trials. *Journal of Biopharmaceutical Statistics* 2004; **14**:263–273.
41. Wang SJ, Hung HMJ. TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine, Special issue: Non-inferiority Trials* 2003; **22**:227–238.
42. Datta S, Halloran ME, Longini IM. Augmented HIV vaccine trial design for estimating reduction in infectiousness and protective efficacy. *Statistics in Medicine* 1998; **17**:185–200.
43. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal* 2001; **25**:435–449.
44. Julious SA, Zariffa N. The ABC of pharmaceutical trial design: some basic principles. *Pharmaceutical Statistics* 2002; **1**:45–53.
45. FDA. Points to consider. Clinical evaluation of anti-infective drug products, 1992.
46. CPMP Notes for guidance on the evaluation of medicinal products indicated for the treatment of bacterial infections, 2004. Doc CPMP/EWP/558/95.
47. Dunnett CW, Gent M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics* 1977; **33**:593–602.
48. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* 1990; **9**:1447–1454.
49. Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985; **4**:213–226.
50. Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 1984; **40**:513–517.

51. Julious SA, Owen RA. Comparison of methods for sample size estimation for non-inferiority studies with binary outcomes. *Statistical Methods in Medical Research* 2011; **20**(6):595–612.
52. Newcombe RG. Interval estimation for the difference between independent proportions:comparison of eleven methods. *Statistics in Medicine* 1998; **17**:873–890.
53. Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* 1995; **5**(3):297–306.
54. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; **11**:283–319.
55. Day S. Clinical trial numbers and confidence intervals of pre-specified size. *The Lancet* 1988; **Dec 17:2**(8625):1427.
56. Wood J, Lambert M. Sample-size calculations for trials in health services research. *Journal of Health Services and Research and Policy* 1999; **4**:226–229.
57. Julious SA, Patterson SD. Sample sizes for estimation in clinical research. *Pharmaceutical Statistics* 2004; **3**:213–215.