Communications for Statistical Applications and Methods
2015, Vol. 22, No. 4, 389–399

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.4.389
Print ISSN 2287-7843 / Online ISSN 2383-4757

Sample Size Calculations for the Development of Biosimilar Products Based on Binary Endpoints

Seung-Ho Kang[1,a], Ji-Yong Jung[a], Seon-Hye Baik[a]

[a]Department of Applied Statistics, Yonsei University, Korea

Abstract

It is important not to overcalculate sample sizes for clinical trials due to economic, ethical, and scientific reasons. Kang and Kim (2014) investigated the accuracy of a well-known sample size calculation formula based on the approximate power for continuous endpoints in equivalence trials, which has been widely used for Development of Biosimilar Products. They concluded that this formula is overly conservative and that sample size should be calculated based on an exact power. This paper extends these results to binary endpoints for three popular metrics: the risk difference, the log of the relative risk, and the log of the odds ratio. We conclude that the sample size formulae based on the approximate power for binary endpoints in equivalence trials are overly conservative. In many cases, sample sizes to achieve 80% power based on approximate powers have 90% exact power. We propose that sample size should be computed numerically based on the exact power.

Keywords: equivalence trial, power, sample size formula, follow-on biologics

1. Introduction

Many best-selling biological products are set to lose their patents over the next few years; constantly, the assessment of biological product biosimilarity for regulatory approval has received significant attention (Chow, 2014; Chow and Liu, 2010; Chow *et al.*, 2009; Chow *et al.*, 2013; Hsieh *et al.*, 2013; Kang and Chow, 2013; Li *et al.*, 2013; US FDA, 2012; World Health Organization, 2009). It is therefore necessary to demonstrate similar qualities, efficacy, and safety for biosimilar products and renovator biological products in order to obtain regulatory approval. Consequently, the characterization of both products are examined by comparing physicochemical properties, biological activities, impurities, and stability. Immunogenicity tests, preclinical studies, and clinical trials are also conducted to demonstrate no clinically significant differences in safety and efficacy. A phase III comparative study (often designed as an equivalence study) is an important step in systematic studies.

The sample size calculation for a phase III clinical trial is important due to economic, ethical, and scientific considerations (Altman, 1980; Moher *et al.*, 1994). This paper emphasize the drawbacks of oversized studies that calculate sample size based on approximate powers. An oversized study results in an unnecessary waste of resources with the potential to expose unnecessarily large number of subjects to potentially harmful or ineffective treatments (Altman, 1980). Therefore, it is important to compute the minimal sample size needed to achieve a pre-specified power, such as 80% or 90%.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A200).

[1] Corresponding author: Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea. E-mail: seungho@yonsei.ac.kr

Published 31 July 2015 / journal homepage: http://csam.or.kr

Kang and Kim (2014) investigated the accuracy of a well-known sample size calculation formula for continuous endpoints in equivalence trials. The formula is given as (Chow *et al.*, 2003, p.60)

$$n_T = kn_R, \qquad n_2 = \frac{(z_\alpha + z_{\beta/2})^2\sigma^2}{(\delta - |\mu_T - \mu_R|)^2}\left(1 + \frac{1}{k}\right) \tag{1.1}$$

to test $H_0 : |\mu_T - \mu_R| \geq \delta$ against $H_a : |\mu_T - \mu_R| < \delta$, where z_α is the upper α quartile of the standard normal distribution (for example, $z_{0.05} = 1.645$); μ_T and μ_R represent the population means of primary endpoints for a biosimilar product and a renovator biological product, respectively; σ^2 is the population variance of the primary endpoint; k is the allocation ratio; n_T and n_R are the sample sizes of a biosimilar product group and the renovator biological product group, respectively. Kang and Kim (2014) found that the sample size calculation based on (1.1) is very conservative, requiring unnecessarily large samples.

The primary endpoint is often binary; therefore, it is important to investigate the accuracy of sample size calculation formulae for binary endpoints in equivalence trials. This paper extends the results of Kang and Kim (2014) to binary endpoints for three popular metrics: the risk difference, the log of the relative risk, and the log of the odds ratio.

This paper is organized as follows. Section 2 reviews the hypotheses of equivalence trials for binary endpoints. Section 3 provides the sample size calculation formulae based on the approximate and exact powers. Section 4 numerically compares approximate powers with exact powers. Section 5 presents the conclusions.

2. Equivalence Trials for Binary Endpoints

Let X_T and X_R denote the number of events of interest from the biosimilar product group and the renovator biological product group, respectively. It is assumed that X_T and X_R follow binomial distributions $B(n_T, p_T)$ and $B(n_R, p_R)$, respectively. There are three popular metrics that can be used to assess the treatment effect estimated from an equivalence trial. The first is the risk difference, $RD = p_T - p_R$, which is the difference between the test and control groups in proportions of outcomes. The second is the relative risk, or risk ratio ($RR = p_T/p_R$), which is the ratio of the rates of unfavorable events in the test and control groups. The third is the odds ratio, which is the ratio of the odds of success (or failure) of the test product relative to the control product. The characteristics of these three metrics are shown in Sinclair and Bracken (1994) and Walter (2000). In this paper, the log of the relative risk and the log of the odds ratio are investigated instead of the relative risk and the odds ratio, as the former allow metrics that are normally distributed and easier to evaluate in the analysis.

The hypotheses of equivalence for the three metrics are given as follows. For the risk difference, it is given as

$$H_0^D : |p_T - p_R| \geq \delta \quad \text{vs.} \quad H_a^D : |p_T - p_R| < \delta, \tag{2.1}$$

where $\delta(> 0)$ is a pre-specified equivalence margin. For the log of the relative risk, it is given as

$$H_0^R : \left|\log\left(\frac{p_T}{p_R}\right)\right| \geq \delta \quad \text{vs.} \quad H_a^R : \left|\log\left(\frac{p_T}{p_R}\right)\right| < \delta. \tag{2.2}$$

For the log of the odds ratio, it is given as

$$H_0^O : \left|\log\left(\frac{p_T/(1-p_T)}{p_R(1-p_R)}\right)\right| \geq \delta \quad \text{vs.} \quad H_a^O : \left|\log\left(\frac{p_T/(1-p_T)}{p_R(1-p_R)}\right)\right| < \delta. \tag{2.3}$$

The hypotheses in (2.1), (2.2), and (2.3) can be decomposed into two one-sided hypotheses. Specifically, the hypothesis in (2.1) can be re-expressed into two one-sided hypotheses as:

$$H_{01}^D : p_T - p_R \leq -\delta \quad \text{vs.} \quad H_{a1}^D : p_T - p_R > -\delta$$

and

$$H_{02}^D : p_T - p_R \geq \delta \quad \text{vs.} \quad H_{a2}^D : p_T - p_R < \delta.$$

The hypotheses in (2.2) and (2.3) can also be decomposed into two one-sided hypotheses. For the log of the relative risk, they are given as

$$H_{01}^R : \log(p_T) - \log(p_R) \leq -\delta \quad \text{vs.} \quad H_{a1}^R : \log(p_T) - \log(p_R) > -\delta$$

and

$$H_{02}^R : \log(p_T) - \log(p_R) \geq \delta \quad \text{vs.} \quad H_{a2}^R : \log(p_T) - \log(p_R) < \delta.$$

For the log of the odds ratio, they are given as

$$H_{01}^O : \log\left(\frac{p_T}{1-p_T}\right) - \log\left(\frac{p_R}{1-p_R}\right) \leq -\delta \quad \text{vs.} \quad H_{a1}^O : \log\left(\frac{p_T}{1-p_T}\right) - \log\left(\frac{p_R}{1-p_R}\right) > -\delta$$

and

$$H_{02}^O : \log\left(\frac{p_T}{1-p_T}\right) - \log\left(\frac{p_R}{1-p_R}\right) \geq \delta \quad \text{vs.} \quad H_{a2}^O : \log\left(\frac{p_T}{1-p_T}\right) - \log\left(\frac{p_R}{1-p_R}\right) < \delta.$$

If two null hypotheses (H_{01}^D and H_{02}^D for the risk difference, H_{01}^R and H_{02}^R for the log of the relative risk, H_{01}^O and H_{02}^O for the log of the odds ratio) in two one-sided hypotheses for each metric are rejected at the significance level α, it can be concluded that the original null hypothesis for each metric (H_0^D for the risk difference, H_0^R for the log of the relative risk, H_0^O for the log of the odds ratio) can be rejected at significance level α. The biosimilar product and the renovator biological product in each case are claimed to be biosimilar.

3. Sample Size Calculation Based on the Approximate and Exact Powers

Section 2 introduces two one-sided hypotheses for the risk difference, the log of the relative risk, and the log of the odds ratio. Test statistics for each hypothesis can be constructed using the central limit theorem, Slutsky's theorem, and the delta method. For the risk difference, the test statistics are given by

$$Z_L^D = \frac{\hat{p}_T - \hat{p}_R - \delta}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_R(1-\hat{p}_R)}{n_R}}}, \qquad Z_U^D = \frac{\hat{p}_T - \hat{p}_R + \delta}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_R(1-\hat{p}_R)}{n_R}}}.$$

For the log of the relative risk, the test statistics are given by

$$Z_L^R = \frac{\log(\hat{p}_T) - \log(\hat{p}_R) - \delta}{\sqrt{\frac{1-\hat{p}_T}{n_T\hat{p}_T} + \frac{1-\hat{p}_R}{n_R\hat{p}_R}}}, \qquad Z_U^R = \frac{\log(\hat{p}_T) - \log(\hat{p}_R) + \delta}{\sqrt{\frac{1-\hat{p}_T}{n_T\hat{p}_T} + \frac{1-\hat{p}_R}{n_R\hat{p}_R}}}.$$

For the log of the odds ratio, the test statistics are given by

$$Z_L^O = \frac{\log\left(\frac{\hat{P}_T}{1-\hat{P}_T}\right) - \log\left(\frac{\hat{P}_R}{1-\hat{P}_R}\right) - \delta}{\sqrt{\frac{1}{n_T\hat{P}_T(1-\hat{P}_T)} + \frac{1}{n_R\hat{P}_R(1-\hat{P}_R)}}}, \qquad Z_U^O = \frac{\log\left(\frac{\hat{P}_T}{1-\hat{P}_T}\right) - \log\left(\frac{\hat{P}_R}{1-\hat{P}_R}\right) + \delta}{\sqrt{\frac{1}{n_T\hat{P}_T(1-\hat{P}_T)} + \frac{1}{n_R\hat{P}_R(1-\hat{P}_R)}}}.$$

For the risk difference, both H_{01}^D and H_{02}^D are rejected at the significance level α if $Z_L^D < -z_\alpha$ and $Z_U^D > z_\alpha$. Similar conclusions can be drawn for the log of the relative risk and the log of the odds ratio using (Z_L^R, Z_U^R) and (Z_L^O, Z_U^O), respectively.

Kang and Kim (2014) showed that, under the alternative hypothesis $H_a : |p_T - p_R| < \delta$, the power of the test for the risk difference is given by

$$\begin{aligned}
&P\left(Z_L^D < -z_\alpha \text{ and } Z_U^D > z_\alpha | H_a\right) && (3.1)\\
&= P\left(Z_L^D < -z_\alpha | H_a\right) + P\left(Z_U^D > z_\alpha | H_a\right) - P\left(Z_L^D < -z_\alpha \quad \text{or} \quad Z_U^D > z_\alpha | H_a\right)\\
&= P\left(Z_L^D < -z_\alpha | H_a\right) + P\left(Z_U^D > z_\alpha | H_a\right) - \left[1 - P\left(Z_L^D \geq -z_\alpha \quad \text{and} \quad Z_U^D \leq z_\alpha | H_a\right)\right]\\
&\geq P\left(Z_L^D < -z_\alpha | H_a\right) + P\left(Z_U^D > z_\alpha | H_a\right) - 1 && (3.2)\\
&= \Psi\left(\frac{\delta - (p_T - p_R)}{\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}}} - z_\alpha\right) + \Psi\left(\frac{\delta + (p_T - p_R)}{\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}}} - z_\alpha\right) - 1\\
&\geq 2\Psi\left(\frac{\delta - |p_T - p_R|}{\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}}} - z_\alpha\right) - 1, && (3.3)
\end{aligned}$$

where Ψ is the cumulative distribution function of the standard normal distribution. The powers in (3.1) and (3.3) are exact and approximate powers, respectively. An advantage of the approximate power is that a closed form of the sample size calculation can be obtained. The sample size needed to achieve power $1 - \beta$ based on the approximate power can be obtained by solving the following equation.

$$1 - \beta = 2\Psi\left(\frac{\delta - |p_T - p_R|}{\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}}} - z_\alpha\right) - 1.$$

Then we have

$$z_{\beta/2} = \frac{\delta - |p_T - p_R|}{\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}}} - z_\alpha.$$

Therefore, the sample size to achieve power $1-\beta$ based on the approximate power to test the hypothesis in (2.1) is

$$n_T = \left(z_\alpha + z_{\frac{\beta}{2}}\right)^2 \frac{\left[\frac{p_T(1-p_T)}{k} + \frac{p_R(1-p_R)}{1}\right]}{(\delta - |p_T - p_R|)^2}, \qquad n_T = kn_R, \tag{3.4}$$

Table 1: Risk difference: the exact and approximate powers ($\alpha = 0.05$)

$n_1 = n_2$	δ	p_1	p_2	$p_1 - p_2$	Exact	Approx
100	0.1	0.052	0.05	0.002	0.8827	0.8677
	0.1	0.054	0.05	0.004	0.8734	0.8422
	0.1	0.056	0.05	0.006	0.8625	0.8139
	0.1	0.058	0.05	0.008	0.8500	0.7827
	0.1	0.060	0.05	0.010	0.8358	0.7487
100	0.2	0.120	0.1	0.020	0.9919	0.9847
	0.2	0.140	0.1	0.040	0.9672	0.9347
	0.2	0.160	0.1	0.060	0.9049	0.8100
	0.2	0.180	0.1	0.080	0.7930	0.5861
	0.2	0.200	0.1	0.100	0.6388	0.2775
100	0.25	0.220	0.2	0.020	0.9894	0.9812
	0.25	0.240	0.2	0.040	0.9736	0.9481
	0.25	0.260	0.2	0.060	0.9399	0.8802
	0.25	0.280	0.2	0.080	0.8814	0.7629
	0.25	0.300	0.2	0.100	0.7942	0.5884
100	0.25	0.320	0.3	0.020	0.9629	0.9389
	0.25	0.340	0.3	0.040	0.9355	0.8768
	0.25	0.360	0.3	0.060	0.8872	0.7769
	0.25	0.380	0.3	0.080	0.8159	0.6329
	0.25	0.400	0.3	0.100	0.7226	0.4456
100	0.3	0.430	0.4	0.030	0.9862	0.9744
	0.3	0.460	0.4	0.060	0.9630	0.9264
	0.3	0.490	0.4	0.090	0.9123	0.8247
	0.3	0.520	0.4	0.120	0.8232	0.6464
	0.3	0.550	0.4	0.150	0.6927	0.3854

where k is an allocation ratio. The sample size calculation formula in (3.4) can be found in Chow *et al.* (2003, p.89). Similarly, the sample size to test the hypothesis in (2.2) is

$$n_T = \left(z_\alpha + z_{\frac{\beta}{2}}\right)^2 \frac{\left[\frac{1-p_T}{kp_T} + \frac{1-p_R}{p_R}\right]}{\left(\delta - \left|\log\left(p_T/p_R\right)\right|\right)^2}, \qquad n_T = kn_R \tag{3.5}$$

and the sample size to test the hypothesis in (2.3) is

$$n_T = \left(z_\alpha + z_{\frac{\beta}{2}}\right)^2 \frac{\left[\frac{1}{kp_T(1-p_T)} + \frac{1}{p_R(1-p_R)}\right]}{\left(\delta - \left|\log\left(\frac{p_T/(1-p_T)}{p_R/(1-p_R)}\right)\right|\right)^2}, \qquad n_T = kn_R. \tag{3.6}$$

Wang *et al.* (2002) obtained the sample size calculation formula in (3.6).

4. Comparison of the Exact and Approximate Power

The closed forms of the sample size calculation formulae based on the approximate power in equivalence trials for binary endpoints were derived in Section 3 and given by (3.4), (3.5), and (3.6). However, the approximate power might be smaller than the exact power because the two inequalities in (3.2) and (3.3) are used to derive the approximate power. Hence, it is important to compare the exact power obtained from (3.1) and the approximate power calculated from (3.3). Both the exact and approximate powers were calculated numerically with R code based on (3.1) and (3.3) as presented in Tables 1–3 (the R code is available from the authors upon request). In all cases, the exact powers are always greater than the approximate powers.

Table 2: Relative risk: the exact and approximate powers ($\alpha = 0.05$)

$n_1 = n_2$	δ	p_1	p_2	p_1/p_2	Exact	Approx
100	0.8	0.32	0.3	1.067	0.9598	0.9340
	0.8	0.34	0.3	1.133	0.9450	0.8946
	0.8	0.36	0.3	1.200	0.9188	0.8389
	0.8	0.38	0.3	1.267	0.8819	0.7642
	0.8	0.40	0.3	1.333	0.8344	0.6689
100	0.6	0.42	0.4	1.050	0.9308	0.8910
	0.6	0.44	0.4	1.100	0.9114	0.8343
	0.6	0.46	0.4	1.150	0.8769	0.7579
	0.6	0.48	0.4	1.200	0.8293	0.6598
	0.6	0.50	0.4	1.250	0.7697	0.5398
100	0.5	0.52	0.5	1.040	0.9409	0.9066
	0.5	0.54	0.5	1.080	0.9236	0.8568
	0.5	0.56	0.5	1.120	0.8925	0.7882
	0.5	0.58	0.5	1.160	0.8486	0.6983
	0.5	0.60	0.5	1.200	0.7926	0.5854
100	0.4	0.62	0.6	1.033	0.9308	0.8907
	0.4	0.64	0.6	1.067	0.9109	0.8327
	0.4	0.66	0.6	1.100	0.8747	0.7531
	0.4	0.68	0.6	1.133	0.8239	0.6488
	0.4	0.70	0.6	1.167	0.7593	0.5188
100	0.4	0.72	0.7	1.029	0.9922	0.9864
	0.4	0.74	0.7	1.057	0.9877	0.9759
	0.4	0.76	0.7	1.086	0.9792	0.9585
	0.4	0.78	0.7	1.114	0.9653	0.9307
	0.4	0.80	0.7	1.143	0.9441	0.8882

Table 3: Odd ratio: the exact and approximate powers ($\alpha = 0.05$)

$n_1 = n_2$	δ	p_1	p_2	$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$	Exact	Approx
100	1.0	0.42	0.4	1.086	0.9218	0.8776
	1.0	0.43	0.4	1.132	0.9086	0.8406
	1.0	0.44	0.4	1.179	0.8899	0.7956
	1.0	0.45	0.4	1.227	0.8657	0.7419
	1.0	0.46	0.4	1.278	0.8361	0.6791
100	1.0	0.52	0.5	1.083	0.9310	0.8918
	1.0	0.53	0.5	1.128	0.9179	0.8566
	1.0	0.54	0.5	1.174	0.8993	0.8128
	1.0	0.55	0.5	1.222	0.8750	0.7596
	1.0	0.56	0.5	1.273	0.8449	0.6962
100	1.0	0.62	0.6	1.088	0.9167	0.8698
	1.0	0.63	0.6	1.135	0.8999	0.8255
	1.0	0.64	0.6	1.185	0.8762	0.7703
	1.0	0.65	0.6	1.238	0.8453	0.7029
	1.0	0.66	0.6	1.294	0.8071	0.6225
100	1.2	0.73	0.7	1.159	0.9525	0.9131
	1.2	0.74	0.7	1.220	0.9341	0.8735
	1.2	0.75	0.7	1.286	0.9083	0.8201
	1.2	0.76	0.7	1.357	0.8739	0.7500
	1.2	0.77	0.7	1.435	0.8297	0.6608
100	1.4	0.82	0.8	1.139	0.9648	0.9391
	1.4	0.83	0.8	1.221	0.9467	0.8996
	1.4	0.84	0.8	1.313	0.9178	0.8397
	1.4	0.85	0.8	1.417	0.8750	0.7526
	1.4	0.86	0.8	1.536	0.8152	0.6320

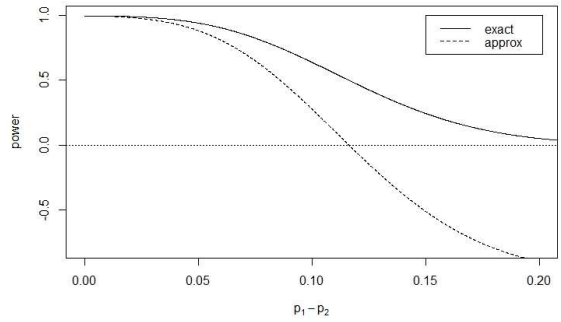

Figure 1: *Comparison of exact and approximate power (risk difference) (p_1 = 0.1–0.3, p_2 = 0.1, δ = 0.2, α = 0.05, $n_1 = n_2$ = 100).*

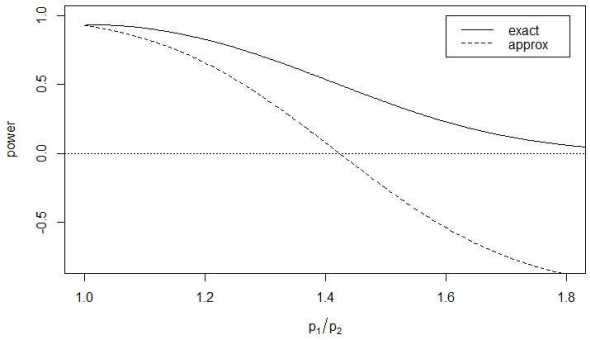

Figure 2: *Comparison of exact and approximate power (relative risk) (p_1 = 0.4–0.8, p_2 = 0.4, δ = 0.6, α = 0.05, $n_1 = n_2$ = 100).*

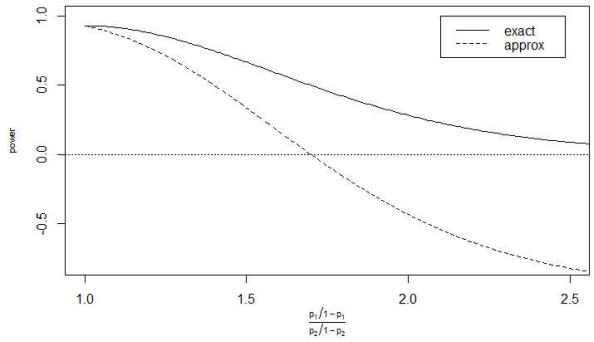

Figure 3: *Comparison of exact and approximate power (odd ratio) (p_1 = 0.4–0.7, p_2 = 0.4, δ = 1, α = 0.05, $n_1 = n_2$ = 100).*

Figure 1 is a graphical representation of the differences between the two powers for the risk difference when α = 5%, $n_1 = n_2$ = 100, and δ = 0.2. As the value of $p_1 - p_2$ increases, the differences between the two curves also increase and means that the accuracy of the approximate power drops rapidly. When the value of $p_1 - p_2$ is greater than 0.12, the approximate power drops below zero, which is unacceptable because powers should be positive. Figures 2 and 3 show similar patterns of

Table 4: Risk difference: sample size calculations based on exact and approximate powers ($\alpha = 0.05$)

δ	p_1	p_2	$p_1 - p_2$	Power	Exact	Approx	Power	Exact	Approx
0.1	0.052	0.05	0.002	80%	84	87	90%	105	110
0.1	0.054	0.05	0.004		85	92		108	116
0.1	0.056	0.05	0.006		88	98		111	123
0.1	0.058	0.05	0.008		90	104		115	131
0.1	0.060	0.05	0.010		93	110		119	139
0.2	0.120	0.1	0.020	80%	44	52	90%	56	66
0.2	0.140	0.1	0.040		54	71		72	89
0.2	0.160	0.1	0.060		72	99		99	124
0.2	0.180	0.1	0.080		103	142		142	179
0.2	0.200	0.1	0.100		155	215		215	271
0.25	0.220	0.2	0.020	80%	47	54	90%	60	68
0.25	0.240	0.2	0.040		52	67		69	85
0.25	0.260	0.2	0.060		62	84		84	106
0.25	0.280	0.2	0.080		78	108		108	136
0.25	0.300	0.2	0.100		102	141		141	178
0.25	0.320	0.3	0.020	80%	61	70	90%	77	88
0.25	0.340	0.3	0.040		66	85		87	107
0.25	0.360	0.3	0.060		77	105		105	133
0.25	0.380	0.3	0.080		96	133		133	167
0.25	0.400	0.3	0.100		124	172		172	217
0.3	0.430	0.4	0.030	80%	48	57	90%	62	73
0.3	0.460	0.4	0.060		55	73		74	92
0.3	0.490	0.4	0.090		69	96		96	121
0.3	0.520	0.4	0.120		94	130		130	164
0.3	0.550	0.4	0.150		134	186		186	235

differences between two powers for the relative risk and the odds ratio.

In order to investigate how many sample size differences are produced by two different powers, the R code was made to compute sample sizes based on exact and approximate powers. Tables 4–6 display sample sizes needed to achieve 80% and 90% power using two different powers for risk difference, the log of the relative risk, and the odds ratio log, respectively. Sample sizes based on approximate powers are greater than those based on exact powers in all investigated cases. For example, when the risk difference is used for $\beta = 0.2$, $\delta = 0.2$, $p_1 = 0.2$, and $p_2 = 0.1$, the sample size based on the approximate power is 215, but the sample size based on the exact power is 155. The two powers produce a difference for 60 patients which may lead to substantial extra costs and ethical concerns.

Kang and Kim (2014) discovered an interesting phenomenon that the sample sizes needed to achieve 80% approximate power are the same as those needed to achieve 90% exact power for a continuous endpoint. Similar phenomena are also observed for a binary endpoint. Such phenomena occur in 34 of 75 cases in Tables 4–6. For example, such an event occurs when $\delta = 0.2$, $p_1 = 0.18$, and $p_2 = 0.1$ in Table 4 ($n_T = n_R = 142$). In Kang and Kim (2014) Theorem 1 for a continuous endpoint explains why such phenomena occur. A similar theorem can be derived for a binary endpoint as follows.

Theorem 1. *Let* $n_T = n_R$ *and*

$$w = z_\alpha - \frac{[p_T - p_R] + \delta}{\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}}}, \quad \textit{for the risk difference},$$

Table 5: Relative risk: sample size calculations based on exact and approximate powers ($\alpha = 0.05$)

δ	p_1	p_2	p_1/p_2	Power	Exact	Approx	Power	Exact	Approx
0.8	0.32	0.3	1.067	80%	62	71	90%	79	90
0.8	0.34	0.3	1.133		64	81		83	102
0.8	0.36	0.3	1.200		69	93		93	117
0.8	0.38	0.3	1.267		78	107		107	136
0.8	0.40	0.3	1.333		91	126		126	159
0.6	0.42	0.4	1.050	80%	71	82	90%	90	103
0.6	0.44	0.4	1.100		73	94		96	118
0.6	0.46	0.4	1.150		80	109		109	137
0.6	0.48	0.4	1.200		92	127		127	161
0.6	0.50	0.4	1.250		109	151		151	191
0.5	0.52	0.5	1.040	80%	68	78	90%	87	99
0.5	0.54	0.5	1.080		70	89		92	112
0.5	0.56	0.5	1.120		76	103		103	130
0.5	0.58	0.5	1.160		87	120		120	151
0.5	0.60	0.5	1.200		103	142		142	179
0.4	0.62	0.6	1.033	80%	71	82	90%	90	103
0.4	0.64	0.6	1.067		74	94		97	119
0.4	0.66	0.6	1.100		81	110		110	138
0.4	0.68	0.6	1.133		94	129		129	163
0.4	0.70	0.6	1.167		113	156		156	197
0.4	0.72	0.7	1.029	80%	45	51	90%	57	64
0.4	0.74	0.7	1.057		46	57		59	72
0.4	0.76	0.7	1.086		48	64		64	80
0.4	0.78	0.7	1.114		53	72		72	91
0.4	0.80	0.7	1.143		60	82		82	104

Table 6: Odd ratio: sample size calculations based on exact and approximate powers ($\alpha = 0.05$)

δ	p_1	p_2	$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$	Power	Exact	Approx	Power	Exact	Approx
1.0	0.42	0.4	1.086	80%	73	85	90%	93	107
1.0	0.43	0.4	1.132		75	92		98	117
1.0	0.44	0.4	1.179		79	101		104	128
1.0	0.45	0.4	1.227		84	112		113	141
1.0	0.46	0.4	1.278		91	124		124	156
1.0	0.52	0.5	1.083	80%	71	82	90%	90	103
1.0	0.53	0.5	1.128		73	89		94	113
1.0	0.54	0.5	1.174		77	98		101	124
1.0	0.55	0.5	1.222		82	108		109	137
1.0	0.56	0.5	1.273		89	120		121	152
1.0	0.62	0.6	1.088	80%	75	86	90%	95	109
1.0	0.63	0.6	1.135		78	95		101	121
1.0	0.64	0.6	1.185		82	106		109	134
1.0	0.65	0.6	1.238		89	119		120	150
1.0	0.66	0.6	1.294		99	135		135	170
1.2	0.73	0.7	1.159	80%	63	77	90%	81	97
1.2	0.74	0.7	1.220		67	86		88	108
1.2	0.75	0.7	1.286		72	97		97	122
1.2	0.76	0.7	1.357		81	110		110	139
1.2	0.77	0.7	1.435		92	127		127	161
1.4	0.82	0.8	1.139	80%	59	70	90%	76	88
1.4	0.83	0.8	1.221		64	80		83	101
1.4	0.84	0.8	1.313		70	93		94	117
1.4	0.85	0.8	1.417		81	110		110	138
1.4	0.86	0.8	1.536		96	133		133	168

Table 7: Further investigation on Theorem 1 ($\alpha = 0.05$)

Metric	δ	p_1	p_2	Sample size with 80% approx power	Sample size with 90% exact power	w
Risk difference	0.2	0.18	0.10	142	142	9.9519×10^{-8}
Risk difference	0.2	0.20	0.10	215	215	4.2490×10^{-13}
Risk difference	0.25	0.28	0.20	108	108	2.4720×10^{-5}
Risk difference	0.25	0.30	0.20	141	141	1.0651×10^{-7}
Relative risk	0.8	0.38	0.30	107	107	9.2356×10^{-5}
Relative risk	0.8	0.40	0.30	126	126	2.2052×10^{-6}
Relative risk	0.6	0.48	0.40	127	127	6.1416×10^{-5}
Relative risk	0.6	0.50	0.40	151	151	1.0050×10^{-6}
Odd ratio	1.2	0.76	0.70	110	110	5.0446×10^{-4}
Odd ratio	1.2	0.77	0.70	127	127	7.0076×10^{-5}
Odd ratio	1.4	0.85	0.80	110	110	5.9863×10^{-4}
Odd ratio	1.4	0.86	0.80	133	133	5.1409×10^{-5}

$$w = z_\alpha - \frac{[\log(p_T) - \log(p_R)] + \delta}{\sqrt{\frac{1-p_T}{n_T p_T} + \frac{1-p_R}{n_R p_R}}}, \quad \textit{for the log of the relative risk,}$$

$$w = z_\alpha - \frac{\log\left(\frac{P_T}{1-P_T}\right) - \log\left(\frac{P_R}{1-P_R}\right) + \delta}{\sqrt{\frac{1}{n_T P_T(1-P_T)} + \frac{1}{n_R P_R(1-P_R)}}}, \quad \textit{for the log of the odds ratio.}$$

When w is so small that $\Psi(w)$ is negligible, the exact power with the sample size to achieve $1-\beta$ approximate power is actually $1-\beta/2$.

Proof: The proof of this theorem is the same as Theorem 1 in Kang and Kim (2014). □

Some cases in which the phenomenon described in Theorem 1 occurs were chosen from Tables 4–6, and the values of w were examined (Table 7). All values of w in Table 7 are small and negligible.

5. Conclusion

In this paper, we studied the accuracy of sample size calculation formulae based on the approximate power for binary endpoints in equivalence trials. The risk difference, the log of the relative risk, and the log of the odds ratio were investigated. Formulae were very conservative because the two inequalities derived the closed form of the sample size calculation based on approximate power. In many practical cases, equivalence trials are planned to achieve 80% power. However, this paper shows that the sample sizes to achieve 80% approximate power often have 90% exact power. Therefore, sample size calculation based on the approximate power may produce unnecessary costs and ethical concerns.

This paper proposes that sample sizes for binary endpoints in equivalence trials should be calculated based on the exact power. The R code to calculate the sample sizes based on the exact power is available from the authors upon request.

References

Altman, D. G. (1980). Statistics and ethics in medical research, III: How large a sample? *The British Medical Journal*, **281**, 1336–1338.

Chow, S. C. (2014). *Biosimilars: Design and Analysis of Follow-on Biologics*, CRC Press, Boca Raton, Florida.

Chow, S. C., Hsieh, T. C., Chi, E. and Yang J. (2009). A comparison of moment-based and probability-based criteria for assessment of follow-on biologics, *Journal of Biopharmaceutical Statistics*, **20**, 31–45.

Chow, S. C. and Liu, J. P. (2010). Statistical assessment of biosimilar products, *Journal of Biopharmaceutical Statistics*, **20**, 10–30.

Chow, S. C., Shao, J. and Wang, H. (2003). *Sample Size Calculations in Clinical Research*, Marcel Dekker, New York.

Chow, S. C., Wang, J., Endrenyi, L. and Lachenbruch, P. A. (2013). Scientific considerations for assessing biosimilar products, *Statistics in Medicine*, **32**, 370–381.

Hsieh, T. C., Chow, S. C., Yang, L. Y. and Chi, E. (2013). The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics, *Statistics in Medicine*, **32**, 406–414.

Kang, S. H. and Chow, S. C. (2013). Statistical assessment of biosimilarity based on relative distance between follow-on biologics, *Statistics in Medicine*, **32**, 382–392.

Kang, S. H. and Kim, Y. (2014). Sample size calculations for the development of biosimilar products. *Journal of Biopharmaceutical Statistics*, **24**, 1215–1224.

Li, Y., Liu, Q., Wood, P. and Johri, A. (2013). Statistical considerations in biosimilar clinical efficacy trials with asymmetrical margins, *Statistics in Medicine*, **32**, 393–405.

Moher, D., Dulberg, C. S. and Wells, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials, *Journal of the American Medical Association*, **272**, 122–124.

Sinclair, J. C. and Bracken, M. B. (1994). Clinically useful measures of effect in binary analyses of randomized trials, *Journal of Clinical Epidemiology*, **47**, 881–889.

US Food and Drug Administration (2012). *Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*, US Food and Drug Administration, Rockville, MD.

Walter, S. D. (2000). Choice of effect measure for epidemiological data, *Journal of Clinical Epidemiology*, **53**, 931–939.

Wang, H., Chow, S. C. and Li, G. (2002). On sample size calculation based on odds ratio in clinical trials, *Journal of Biopharmaceutical Statistics*, **12**, 471–483.

World Health Organization (2009). *Guidelines on Evaluation of Similar Biotherapeutic Products (SBPs)*, World Health Organization, Geneva.

Received June 17, 2015; Revised July 14, 2015; Accepted July 14, 2015