Sample Size Calculation in Clinical Trials with Binary Co-primary Endpoints or Multiple Testing Procedures

Zuoshun Zhang[1]
Biostatistics, Celgene Corporation, Summit, New Jersey, USA

Abstract
For clinical trials with multiple co-primary binary endpoints, we present an efficient method in power and sample size estimation. The method involves simulating sufficient test statistical vectors and can be easily implemented by a statistical practitioner. Its outputs on power and sample size are accurate in comparison with outcomes from numerical integration method or from simulating correlated binary endpoints of individual subjects. When number of endpoints increase, the complexity of the method remains the same. We also illustrate the method for power and sample size estimation using Holm and Hochberg procedures for multiplicity adjustment.

Key Words: sample size, co-primary, binary endpoint, multiple testing procedure, simulation, power

1. Introduction

Clinical trials generally pose multiple questions in the form of hypotheses whose evaluations involve multiple comparisons and tests for multiple endpoints (Dmitrienko et al. 2009). The trials may involve comparisons of one or more active arms with a control arm in several primary endpoints. The win criteria range from statistically significant for at least one comparison to statistically significant for all comparisons in co-primary designs. In the design stage, we need to estimate power and sample size with appropriate multiplicity adjustments, where there may not have readily available software.

In clinical trials with co-primary endpoints, a non-significant result in any one of the specified efficacy endpoints would lead to a non-win scenario for the trial. Examples can be found in migraine, acute pain, psoriasis, Alzheimer's disease, menopausal symptoms, and vaccine trials. Offen et al. (2007) and Dmitrienko et al. (2009) reviewed co-primary designs with medical and statistical perspectives. Several methods for power and sample size calculations have been proposed for clinical trials with multiple co-primary endpoints (Xiong et al., 2005; Eaton et al. 2007; Hung and Wang 2009; Song, 2009; Kordzakhia et al. 2010; Sozu et al. 2010, 2011; Julious et al. 2012). Xiong et al. (2005) proposed a power formula for two co-primary endpoints in superiority clinical trials when the endpoints are bivariate normally distributed and their variance–covariance matrix is known. Eaton et al. (2007) provided computable bounds of the power function under the assumption of multivariate normality. Hung and Wang (2009) gave bounds for sample size. Song (2009) discussed sample size calculations with multiple co-primary binary endpoints for non-inferiority clinical trials. Kordzakhia et al. (2010) presented a generalization in testing co-primary endpoints using a method of balanced adjustment. Sozu et al. (2010, 2011) provided formulas for power and sample size calculations with multiple co-primary continuous and binary endpoints in superiority clinical trials. In addition, Kong et al. (2004) used simulation and presented type I error and power estimation in non-inferiority and equivalence trials with correlated multiple endpoints.

[1] Address correspondence to Zuoshun Zhang, Celgene Corporation, 300 Connell Drive, Berkeley Heights, NJ 07922, USA; E-mail: zzhang@celgene.com

Bang et al. (2005) calculated sample size for simulation-based multiple-testing procedures.

For clinical trials with multiple co-primary continuous endpoints, Sozu et al. (2011) provided formulas for power and sample size calculations in case of one to one randomization. They included tables for sample sizes using numerical integration. Zhang (2012) derived formulas for more general cases and used simulation method to estimate power and sample sizes and showed that the differences between the two methods are very small in all checked cases.

For clinical trials with multiple co-primary binary endpoints, Sozu et al. (2010) provided formulas for power and sample size calculations and illustrated their method for 2 or 3 endpoints using numerical integration. In this paper, we are going to revisit the problem using similar method as in Bang et al (2005). In Section 2, we are going to present power and sample size estimation for co-primary binary endpoints in a more general setting. Section 2.1 presents notations for co-primary binary design. Section 2.2 presents power and sample size formulas by transforming the sufficient test vector for chi-square test into a standard random vector under the alternative hypotheses and derives its correlation matrix. Section 2.3 lists steps to estimate power and sample size using SAS programs. Section 2.4 and 2.5 present numerical examples with $K = 2$ and $K \geq 3$ endpoints and compared with sample sizes in Sozu et al. (2010). For non-negative correlated binary endpoints, Section 2.6 outlines an alternative approach for power estimation by simulating binary outcomes from individual subjects, which results in similar powers. Section 3 generalizes power and sample size estimation for co-primary binary endpoints to more general cases not requiring significant of all primary endpoints. Examples for sample size with Holm and Hochberg procedure are provided. Section 4 puts forth a discussion and conclusion. Sample SAS programs are attached in the appendix.

2. Two Arm Clinical Trials with Multiple Co-Primary Binary Endpoints

In this section, we set up notations for statistical design with co-primary binary endpoints. We use different notations from Sozu et al. (2010) in order to have uniform presentation for overall sample sizes for trials with different ratios of randomization between two arms (Rosenberger and Lachin 2002). Similar approaches were presented in Zhang (2012) for designs with continuous co-primary endpoints. We use the chi-square test for comparisons between endpoints in the two arms. By transforming the test statistic vectors into standard random vectors under the alternative hypotheses, we derive an equation connecting power and sample size, which will be the basis for evaluating power using simulation. Then we list steps to simulate correlated random vectors in power estimation. Finally, we present numerical results for sample size and compare with Sozu et al. (2010).

2.1 Statistical Model

In a randomized clinical trial with two arms, let $i = A$ or C denote the test or the control arms. Overall, the trial randomizes n subjects, where $n_A = Q_A n$ denotes the number of subjects randomized to arm A and $n_C = Q_C n = (1 - Q_A)n$ denotes the number of subjects randomized to arm C. There are $k = 1, \cdots, K$ co-primary binary endpoints with known correlation coefficients.

For a subject in arm A, the response vector $Y_{A,j} = \left(Y_{A1,j}, \cdots, Y_{AK,j}\right)^T$, where $j = 1, \ldots, n_A$; and for a subject in arm C, the response vector $Y_{C,j} = \left(Y_{C1,j}, \cdots, Y_{CK,j}\right)^T$, where $j = 1, \ldots, n_C$ .The variables $Y_{ik,j}$ have value 1 for response and 0 for non-response. All subjects are assessed for all K endpoints. Those endpoints are correlated.

For $i = A, C; k, k' = 1, \cdots, K; j = 1, \cdots, n_i$, we use the following notation for response probabilities and correlation coefficients:

$$E\left(Y_{ik,j}\right) = Prob(\left(Y_{ik,j} = 1\right) = \pi_{ik}$$

$$Corr\left(Y_{ik,j}, Y_{ik',j}\right) = \frac{E\left(Y_{ik,j}Y_{ik',j}\right) - \pi_{ik}\pi_{ik'}}{\sqrt{\pi_{ik}(1-\pi_{ik})\pi_{ik'}(1-\pi_{ik'})}} = \rho_i^{kk'}$$

Since the values for response rates π_{ik} and $\pi_{ik'}$ range from 0 to 1, the correlation coefficients cannot freely take values between -1 and 1. From the above formula, it can be easily derived that the correlation coefficients are bounded from below by:

$$\max\left[-\sqrt{\frac{\pi_{ik}\pi_{ik'}}{(1-\pi_{ik})(1-\pi_{ik'})}}, -\sqrt{\frac{(1-\pi_{ik})(1-\pi_{ik'})}{\pi_{ik}\pi_{ik'}}}\right],$$

and from above by

$$\min\left[\sqrt{\frac{\pi_{ik}(1-\pi_{ik'})}{(1-\pi_{ik})\pi_{ik'}}}, \sqrt{\frac{(1-\pi_{ik})\pi_{ik'}}{\pi_{ik}(1-\pi_{ik'})}}\right].$$

The bounds provide necessary conditions for correlation coefficients between binary endpoints.

We are interested in testing the hypotheses with focus on the differences between the response rates $\delta_{AC}^k = \pi_{Ak} - \pi_{Ck}$, where we assume a positive value of δ_{AC}^k indicates the benefit of the test arm A over the control arm C, $k = 1, \ldots, K$.

For $k = 1, \ldots, K$, the null and alternative hypotheses are expressed as

$$H_{AC,0}^k: \delta_{AC}^k \leq 0 \; vs. H_{AC,1}^k: \delta_{AC}^k > 0.$$

The co-primary design seeks to assert the superiority of the test arm A over the control arm C in all K primary endpoints simultaneously. The trial is designed to reject all K null hypotheses at the same time at the one-sided level α (usually 0.025) with the overall power $\geq 1-\beta$ using the pre-specified response rates π_{Ak} and π_{Ck} and correlation coefficients between two endpoints in both arms $\rho_A^{kk'}$ and $\rho_C^{kk'}$, where $k, k' = 1, \ldots, K$.

2.2 Power and Sample Size Using Chi-square Test

The test statistic for the chi-square test without continuity correction can be conveniently expressed as Z score using normal approximation. For arm A or C, endpoint $k = 1, \ldots, K$, the observed response rates can be expressed as

$$p_{Ak} = \overline{\mathrm{Y}}_{\mathrm{Ak}} = \sum_{j=1}^{n_A} Y_{Ak,j}/n_A, p_{Ck} = \overline{\mathrm{Y}}_{\mathrm{Ck}} = \sum_{j=1}^{n_C} Y_{Ck,j}/n_C.$$

The pooled observed response rates can be expressed as

$$p_{AC}^k = (n_A p_{Ak} + n_C p_{Ck})/n = Q_A p_{Ak} + Q_C p_{Ck}$$

For the null hypothesis$H_{AC,0}^k$, the test statistic formula is

$$Z_{AC}^k = \frac{p_{Ak} - p_{Ck}}{\sqrt{var(p_{Ak} - p_{Ck}|H_0)}} = \frac{p_{Ak} - p_{Ck}}{\emptyset_{AC,0}^k/\sqrt{n}}$$

where the pooled variance under the null hypothesis is expressed as

$$\emptyset_{AC,0}^k = \sqrt{\pi_{AC}^k(1 - \pi_{AC}^k)(1/Q_A + 1/Q_C)}$$

with

$$\pi_{AC}^k = Q_A \pi_{Ak} + Q_C \pi_{Ck}$$

The test statistic vector $Z_{AC} = (Z_{AC}^1, \cdots, Z_{AC}^K)^T$ follows multi-normal distribution asymptotically. Its covariance matrix can be derived when the correlation coefficients among the K binary endpoints are known.

.

The co-primary design rejects all K null hypotheses when all $Z_{AC}^k \geq z_\alpha, k = 1, \ldots, K$, with the one-sided test at the significant level α, where z_α is the upper 100αth percentile of the standard normal distribution. The power function is evaluated under the alternative hypotheses at pre-specified response rates and correlation coefficients π_{Ak}, π_{Ck}, $\rho_A^{kk'}$, $\rho_C^{kk'}$, for $k, k' = 1, \ldots, K$. The overall power can be expressed as

$$Power = Prob\left[\bigcap_{k=1}^{K}\{Z_{AC}^k \geq z_\alpha|H_1\}\right] = Prob\left[\bigcap_{k=1}^{K}\{W_{AC}^k \geq w_{AC}^k(\alpha, n)|H_1\}\right],$$

where the transformed test statistic vector $W_{AC} = (W, \cdots, W_{AC}^K)^T$ follows multivariate normal distribution asymptotically with mean vector 0, variance 1, and known correlation coefficients. The formula is

$$W_{AC}^k = \frac{\emptyset_{AC,0}^k}{\emptyset_{AC,1}^k} Z_{AC}^k - \frac{\delta_{AC}^k}{\emptyset_{AC,1}^k}\sqrt{n}$$

where

$$\emptyset_{AC,1}^k = \sqrt{\frac{\pi_{Ak}(1 - \pi_{Ak})}{Q_A} + \frac{\pi_{Ck}(1 - \pi_{Ck})}{Q_C}}$$

and the constants are given as

$$w_{AC}^k(\alpha, n) = \frac{\emptyset_{AC,0}^k}{\emptyset_{AC,1}^k} z_\alpha - \frac{\delta_{AC}^k}{\emptyset_{AC,1}^k}\sqrt{n}$$

The correlation coefficients can be easily derived as

$$r_{AC}^{kk'} = corr\left(W_{AC}^k, W_{AC}^{k'}\right)$$

$$= E\left[\frac{p_{Ak} - p_{Ck} - \delta_{AC}^k}{\emptyset_{AC,1}^k/\sqrt{n}} \cdot \frac{p_{Ak'} - p_{Ck'} - \delta_{AC}^{k'}}{\emptyset_{AC,1}^{k'}/\sqrt{n}}\right]$$

$$= \frac{\rho_A^{kk'}\emptyset_A^k\emptyset_A^{k'}/Q_A + \rho_C^{kk'}\emptyset_C^k\emptyset_C^{k'}/Q_C}{\emptyset_{AC,1}^k\emptyset_{AC,1}^{k'}}$$

where

$$\emptyset_i^k = \sqrt{\pi_{ik}(1 - \pi_{ik})},\ i = A, B; k = 1, \cdots K.$$

We denote the correlation matrix by

$$R_{AC} = \begin{pmatrix} 1 & \cdots & r_{AC}^{1K} \\ \vdots & \ddots & \vdots \\ r_{AC}^{1K} & \cdots & 1 \end{pmatrix}.$$

Sozu et al. (2010) calculated sample sizes for cases with 2 and 3 ($K = 2, 3$) co-primary endpoints using numerical integration in the power equation. In this note, we simulate the transformed test statistic vector $W_{AC} = (W, \cdots, W_{AC}^{K})^{T}$ to estimate power and sample size. The procedures and programs are essentially the same for any number of co-primary endpoints.

2.3 Power and Sample Size Estimation by Simulating Test Statistics

For each of the $k = 1, \ldots, K$ endpoints, calculate sample sizes for designs using the one-sided test at level α and with power: (i) $1 - \beta$, or (ii) $1 - \beta_1 = \sqrt[K]{(1 - \beta)}$. The overall sample size for designs with K co-primary endpoints would be approximately ranged between the smallest sample size in (i) and the largest sample size in (ii). In the following, we list the steps for power and sample size estimation via simulation of the transformed test statistic vector:

(1) Using SAS PROC IML, check whether the correlation matrix R_{AC} for the transformed test statistic vector $W_{AC} = (W_{AC}^{1}, \cdots, W_{AC}^{K})^{T}$ is positive-definite by showing all K eigenvalues are positive. Otherwise adjust the correlation coefficients between the binary endpoints to make R_{AC} positive-definite.

(2) Let the number of the simulated trials be M (we will use $M = 50000$ in all examples). Use SAS PROC IML to generate M independent K-dimensional vectors of random numbers $W_{AC,m} = \left(W_{AC,m}^{1}, \cdots, W_{AC,m}^{K}\right)^{T}$ from normal distribution with means 0 and variance-covariance matrix R_{AC}. Each $W_{AC,m}$ represents a transformed outcome statistic for a simulated trial under the alternative hypotheses. Specifically, for $m = 1, \cdots, M$, the components of the independently distributed random vectors $W_{AC,m}$ satisfy

$$E\left(W_{AC,m}^{k}\right) = 0,$$
$$Var\left(W_{AC,m}^{k}\right) = 1,$$
$$corr\left(W_{AC,m}^{k}, W_{AC,m}^{k'}\right) = r_{AC}^{kk'}.$$

(3) Let n be the sample size for the trial. For $m = 1, \cdots, M$, find cases which lead to rejection of all co-primary null hypotheses at the one-sided level α, ie, those that satisfy the following condition

$$\left\{W_{AC,m}^{k} \geq w_{AC}^{k}(\alpha, n) : \; k = 1, \cdots, K\right\}.$$

The simulated power is the proportion of all such vectors among the M vectors of random numbers.

(4) For a design to achieve power $(1 - \beta)$, the sample size is estimated as the smallest integer n such that the associated estimated power is greater than or equal to $(1 - \beta)$.

2.4 Numerical Examples for $K = 2$ Co-primary Endpoints

In the following, the sample size estimation is based on the chi-square test at the one-sided level $\alpha = 0.025$.

Sozu et al. (2010) Table III presented sample sizes for 34 cases of different combinations of response rates and correlation coefficients for $K = 2$ co-primary binary endpoints. For the first 21 cases using the chi-square test with sample sizes ranging from 145 to1142, we estimate sample size using simulation method and present outputs in Table 2.1. In all 21 cases, the differences between their results using numerical integration and our results using simulation for sufficient test statistic vectors are within ± 5 and $\pm 1.3\%$. In most cases, there is no difference or the difference is ± 1.

Table 2.1: Sample Size with $K = 2$ Co-primary Binary Endpoints, 1:1 Randomization, One-sided Level $\alpha = 0.025$ and Power $1 - \beta = 0.80$ (80%)

Primary Endpoint	Response Rate		Sample Size for Individual Endpoint with Power		Correlation Coefficients	Sample Size
k	π_{Ak}	π_{Ck}	$1-\beta$	$\sqrt[K]{(1-\beta)}$	$\rho_A^{12} = \rho_C^{12}$	n
1	0.70	0.50	188	244	-0.3	247
2	0.70	0.50	188	244	0.0	244
					0.3	239
					0.5	233
					0.8	218
1	0.87	0.70	182	238	0.0	241
2	0.70	0.50	188	244	0.3	235
					0.5	230
1	0.90	0.70	124	160	0.0	162
2	0.90	0.70	124	160	0.3	158
					0.5	154
					0.8	145
1	0.95	0.90	870	1142	0.0	1142
2	0.95	0.90	870	1142	0.3	1116
					0.5	1089
					0.8	1019

In Phase 3 clinical trials, studies can be designed to have higher power and more more subjects in the active treatment arm. The targeted response rates for the co-primary endpoints may not be the same in one or both arms. The correlation coefficients between endpoints may be different for the active and control arms under the alternative hypothesis. We present examples for sample size estimation in the next table using simulation approach.

Table 2.2: Sample Size with K = 2 Co-primary Binary Endpoints, 2:1 Randomization, One-sided Level $\alpha = 0.025$ and Power $1 - \beta = 0.90$ (90%)

Primary Endpoint	Response Rate		Sample Size for Individual Endpoint with Power		Correlation Coefficients		Sample Size
k	π_{Ak}	π_{Ck}	$1-\beta$	$\sqrt[K]{(1-\beta)}$	ρ_A^{12}	ρ_C^{12}	n
1	0.30	0.10	188	228	0.0	0.0	227
2	0.30	0.10	188	228	0.3	0.3	225
					0.5	0.5	222
					0.7	0.3	221
					0.7	0.7	216
					0.95	0.95	201
					0.999	0.999	191
1	0.30	0.10	188	228	0.0	0.0	252
2	0.25	0.08	225	273	0.3	0.3	250
					0.5	0.5	246
					0.7	0.3	246
					0.7	0.7	242

The following graph power over sample size for different correlation coefficients.

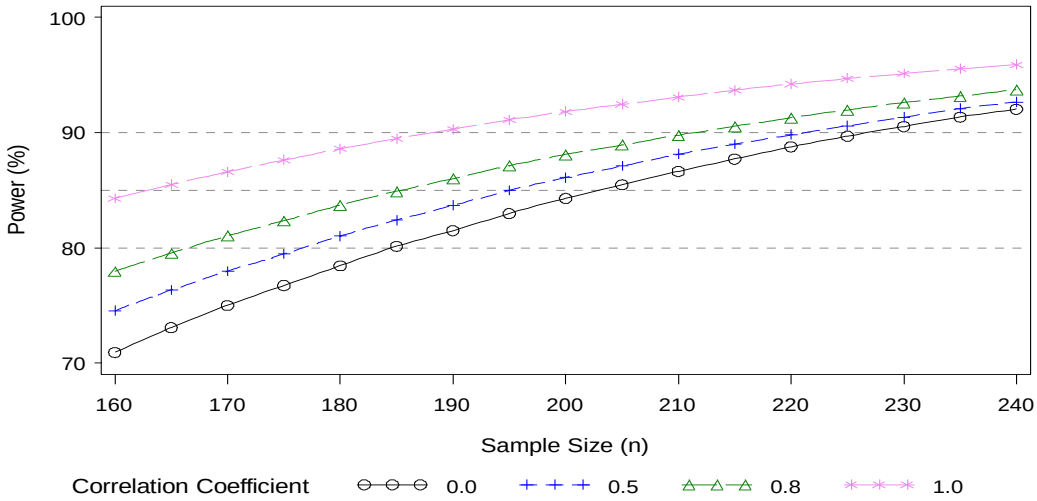

Figure 2.1: Power and Sample Size Using Chi-square Test for $K = 2$ Co-primary Binary Endpoints, 2:1 Randomization, One-sided Level $\alpha = 0.025$, Equal Correlation Coefficients in Both Arms, Response Rates $\pi_{A1} = \pi_{A2} = 0.30, \pi_{C1} = \pi_{C2} = 0.10$

2.5 Numerical Examples for $K \geq 3$ Co-primary Endpoints

Sozu et al. (2010) Table IV presented sample sizes for 24 cases for $K = 3$ co-primary binary endpoints. For the first 14 cases with chi-square test, we estimate sample sizes

using the simulation approach. In all the 14 cases, the differences between their results and our results are within ± 2 and $\pm 1\%$.

Table 2.2: Sample Size with $K = 3$ Co-primary Binary Endpoints, 1:1 Randomization, One-sided Level $\alpha = 0.025$ and Power $1 - \beta = 0.80$ (80%)

Primary Endpoint	Response Rate		Sample Size for Individual Endpoint with Power		Correlation Coefficients	Total Sample Size
k	π_{Ak}	π_{Ck}	$1-\beta$	$\sqrt[K]{(1-\beta)}$	$(\rho_A^{12}, \rho_A^{13}, \rho_A^{23}) = (\rho_C^{12}, \rho_C^{13}, \rho_C^{23})$	n
1	0.70	0.50	188	276	(-0.3, -0.3, 0.0)	281
2	0.70	0.50	188	276	(-0.3, -0.3, 0.3)	278
3	0.70	0.50	188	276	(-0.3, -0.3, 0.5)	274
					(-0.3, -0.3, 0.8)	266
					(0.0, 0.0, 0.0)	278
					(0.0, 0.0, 0.3)	274
					(0.0, 0.0, 0.5)	270
					(0.0, 0.0, 0.8)	262
					(0.3, 0.3, 0.3)	268
					(0.3, 0.3, 0.5)	264
					(0.3, 0.3, 0.8)	256
					(0.5, 0.5, 0.5)	258
					(0.5, 0.5, 0.8)	250
					(0.8, 0.8, 0.8)	234

In general when number of co-primary endpoints increases, power will decrease. The following graph plots one such case with $K = 1, 2, 3, 4$ co-primary endpoints.

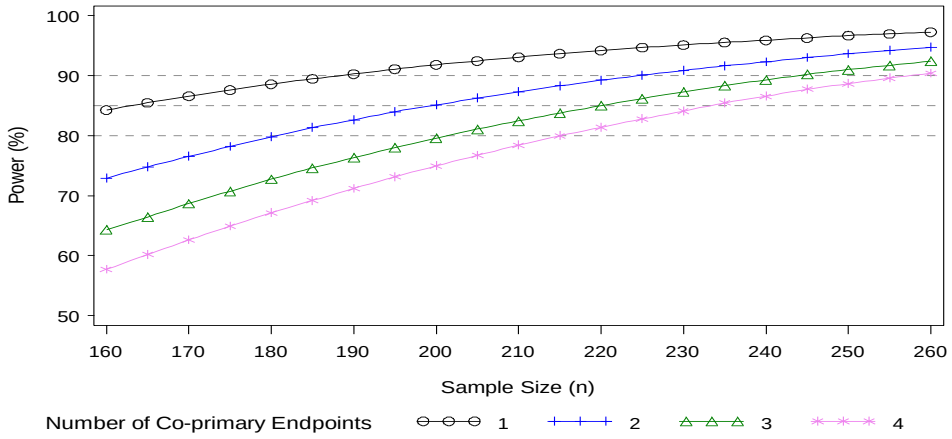

Figure 2.2: Power and Sample Size Using Chi-square Test for Co-primary Designs, 2:1 Randomization, One-sided Level $\alpha = 0.025$, Response Rates 0.30 in Active and 0.10 in Control Arms, Correlation Coefficients 0.30 between Two Endpoints in Both Arms

2.6 Power Estimation by Simulating Correlated Binary Outcomes from Individual Subjects

When sample size is fixed, an alternative way to estimate power is to simulate correlated binary outcomes from individual subjects. Park et al (1996) gave a simple way to represent non-negative correlated binary variables using correlated Poisson variables. Park et al presented formulas for deriving Poisson parameters for two binary endpoints and gave an algorithm for three or more endpoints. In most clinical trials, since the binary endpoints are positively correlated, we can use their method for simulation.

In the following, we are going to restrict to $K = 2$ co-primary binary endpoints. A similar approach applies to any number of binary endpoints with non-negatively correlation coefficients. We are using the same notations as in the previous sections. For arms $i = A, C$ respectively, let the response rates be π_{i1} and π_{i2} for the two endpoints, their correlation coefficient ρ_i^{12}, and sample size $n = n_A + n_C$. In the following, we list steps for power estimation by simulating correlated binary outcomes of individual subjects:

(1) Calculate constants used in simulating Poisson random variables:

$$\alpha_i^{11} = -\log(\pi_{i1}),$$
$$\alpha_i^{22} = -\log(\pi_{i2}),$$
$$\alpha_i^{12} = \log\left(1 + \rho_i^{12}\sqrt{(1-\pi_{i1})(1-\pi_{i2})/(\pi_{i1}\pi_{i2})}\right)$$

(2) Let the number of the simulated trials be M (we use $M = 10000$ in all examples). In one SAS data step, we output $n * M$ three-dimensional independent Poisson random vectors. For arms $i = A$ or C, the 3 components of each vector come from independent Poisson distribution $X_{i1}(\alpha_i^{11} - \alpha_i^{12})$, $X_{i2}(\alpha_i^{22} - \alpha_i^{12})$, and $X_{i3}(\alpha_i^{12})$ with means $(\alpha_i^{11} - \alpha_i^{12})$, $(\alpha_i^{22} - \alpha_i^{12})$, and α_i^{12}, respectively. One simulated trial uses n_A vectors for arm A and n_C vectors for arm C.

(3) For arms $i = A$ or C, define binary random variables as $Y_{i1} = I_{\{0\}}(X_{i1} + X_{i3})$, and $Y_{i2} = I_{\{0\}}(X_{i2} + X_{i3})$, where $I_{\{0\}}(\cdot)$ is the indicating function at integer 0. The binary variables Y_{i1} and Y_{i1} have response rates π_{i1}, π_{i2}, and correlation coefficient ρ_i^{12}.

(4) For each simulated trial with $n_A$vectors for arm A and $n_C$vectors for arm C, use SAS proc freq to calculate p-value from chi-square test.

(5) Count number of simulated trials having one-sided p-values $\leq \alpha$ for both endpoints. The estimated power for the co-primary design is the percentage of those significant trials among the total M trials.

We use the above approach to estimate power using sample sizes in Table 2.1 and 2.2. The estimated powers are presented in the following table for the first 7 and the first 5 cases with non-negative correlation coefficients in Table 2.1 and Table 2.2, respectively. In all cases we checked, the estimated power from simulating correlated binary outcomes from individual subjects are near the target power (80% or 90%) used in sample size estimation using sufficient statistic vectors.

Table 2.5: Power Estimation with $K = 2$ Co-primary Binary Endpoints Using One-sided Chi-square Test at Level $\alpha = 0.025$ for Selected Cases in Table 2.1 and 2.2

Primary Endpoint	Response Rate		Correlation Coefficients		Sample Size			Estimated Power [a]
k	π_{Ak}	π_{Ck}	ρ_A^{12}	ρ_C^{12}	n	n_A	n_B	%
1	0.70	0.50	0.0	0.0	244	122	122	79.97
2	0.70	0.50	0.3	0.3	239	120	119	80.06
			0.5	0.5	233	117	116	79.61
			0.8	0.8	218	109	109	79.60
1	0.87	0.70	0.0	0.0	241	121	120	80.03
2	0.70	0.50	0.3	0.3	235	118	117	80.66
			0.5	0.5	230	115	115	81.17
1	0.30	0.10	0.0	0.0	227	151	76	90.51
2	0.30	0.10	0.3	0.3	225	150	75	90.18
			0.5	0.5	222	148	74	90.58
			0.7	0.3	221	147	74	90.29
			0.7	0.7	216	144	72	90.31

Note [a]: Percentage of trials with significant outcomes for both endpoints using chi-square test at one-sided level $\alpha = 0.025$ among $M = 10000$ simulated trials.

3. Power and Sample Size Estimation Using Holm and Hochberg Procedures for Multiplicity Adjustment

In designs with multiple primary endpoints and the win scenario does not necessarily require statistically significant in all of them, one may need procedures for multiplicity adjustment in order to control family-wise Type I error rate. For those procedures based on univariate p-values, sample size and power can be conveniently estimated using the same approach as in Section 2 for the co-primary designs. In Section 3, we are going to present results for correlated binary endpoints using Holm and Hochberg procedures for trials with one active arm and one control arm using binary endpoints. The same approach applies to more general procedures and designs with more than one active arm versus one control arm.

For Holm and Hochberg procedures, the win scenario is to claim significant in at least one primary endpoint. During the study design stage, we generally evaluate: (i) power for claiming significance of at least one endpoint, and (ii) power for claiming significance in $1 < K' \leq K$ endpoints. Sample sizes can be estimated to have power $\geq 1 - \beta$ using either power definitions (i) or (ii) at the pre-specified alternative response rates π_{Ak} and π_{Ck} and correlation coefficients between two endpoints $\rho_A^{kk'}$ and $\rho_C^{kk'}$, where $k, k' = 1, \ldots, K$.

For easy presentation, we are going to restrict to $K = 2$ binary endpoints and use the same notations as in Section 2. Let the two null hypotheses be $H_{AC,0}^1$ and $H_{AC,0}^2$ for comparisons of the two primary endpoints between the active and the control arms and

the corresponding one-sided p-values be p_{AC}^1 and p_{AC}^2. Denote $p_{(1)} = \min(p_{AC}^1, p_{AC}^2)$ and $p_{(2)} = \max(p_{AC}^1, p_{AC}^2)$ for the most and the least significant comparisons.

The Holm procedure involves the following two steps:

- Step 1. If $p_{(1)} \leq \alpha/2$, reject the corresponding null hypothesis and go to the next step. Otherwise retain both null hypotheses and stop.
- Step 2. If $p_{(2)} \leq \alpha$, reject the corresponding null hypothesis. Otherwise retain the hypothesis and stop.

The Hochberg procedure involves the following two steps:

- Step 1. If $p_{(2)} \leq \alpha$, reject both null hypotheses and stop. Otherwise retain the corresponding null hypothesis and go to the next step.
- Step 2. If $p_{(1)} \leq \alpha/2$, reject the corresponding hypothesis. Otherwise retain the hypothesis and stop.

The win scenarios are to claim significant in at least one endpoint. During study design stage, there is interest to evaluate two different kinds of powers: (i) power for claiming significance of at least one endpoint, and (ii) power for claiming significance of both endpoints. Sample sizes can be estimated to have power $\geq 1-\beta$ using either power definitions (i) or (ii) at the pre-specified alternative response rates π_{Ak} and π_{Ck} and correlation coefficients between two endpoints $\rho_A^{kk'}$ and $\rho_C^{kk'}$, where $k, k' = 1, \ldots, K$.

For endpoint $k = 1, 2$, define critical regions at either α or $\alpha/2$ level as the following:

$$R_\alpha^k = \{Z_{AC}^k \geq z_\alpha\} = \{W_{AC}^k \geq w_{AC}^k(\alpha, n)\},$$
$$R_{\alpha/2}^k = \{Z_{AC}^k \geq z_{\alpha/2}\} = \{W_{AC}^k \geq w_{AC}^k(\alpha/2, n)\},$$

where z_α is the upper 100αth percentile of the standard normal distribution

The power for claiming significance of at least one endpoint can be expressed as

$$Power1_{Holm} = Prob\left[R_{\alpha/2}^1 \cup R_{\alpha/2}^2 | H_1\right],$$

and

$$Power1_{Hochberg} = Prob\left[R_{\alpha/2}^1 \cup R_{\alpha/2}^2 \cup (R_\alpha^1 \cap R_\alpha^2) | H_1\right],$$

for Holm and Hochberg procedures, respectively.

The power for claiming significance of both endpoints can be expressed as

$$Power2_{Holm} = Prob\left[(R_{\alpha/2}^1 \cap R_\alpha^2) \cup (R_\alpha^1 \cap R_{\alpha/2}^2) | H_1\right],$$

and

$$Power2_{Hochberg} = Prob\left[R_\alpha^1 \cap R_\alpha^2 | H_1\right],$$

for Holm and Hochberg procedures, respectively.

Similar steps as those for co-primary designs in Section 2.3 can be used to estimate power and sample size for Holm and Hochberg procedures.

4. Discussion and Conclusion

For design of clinical trials with co-primary binary endpoints, we show that power and sample size estimation can be accurately and efficiently carried out through simulation of sufficient test statistic vectors. For cases with 2 or 3 binary co-primary endpoints, the

differences are small in sample sizes compared to those using numerical integration. In addition, differences in power are small compared to those using simulation of correlated binary outputs from individual subjects. In the appendix, we provide a SAS program for power and sample size estimation for two co-primary binary endpoints by simulating sufficient test statistic vector. The methods can be easily adapted to other cases with multiple correlated binary endpoints such as those using Holm or Hochberg procedures for multiplicity adjustment. The complexity of the simulation approach is essentially the same for different numbers of correlated binary endpoints, and the calculating time won't increase dramatically with large sample sizes. In practice, we recommend estimating power and sample size by simulating sufficient test statistic vectors. In addition, one may verify the power by simulating individual outputs with non-negative correlation coefficients. Similar approach can also be applied to designs with more than one active arms and a control arm.

Acknowledgments

The author is grateful to Xiaolong Luo and Huiling Li for insightful discussion and Gary Cline for comments.

References

Bang, H., Jung, S-H, George, SL (2005). Sample size calculation for simulation-based multiple-testing procedures. J Biopharm Stat. 15(6):957–967.

Berger RL (1982). Multiparameter hypothesis testing and acceptance sampling. Technometrics, 24:295-300.

Chuang-Stein, C, Stryszak, P, Dmitrienko, A and Offen, W (2007), Challenge of multiple co-primary endpoints: a new approach. Statist. Med. 26: 1181–1192.

Dmitrienko, A, Tamhane, AC, and Bretz, F (eds) (2009). Multiple Testing Problems in Pharmaceutical Statistics. CRC Press, New York.

Eaton ML, Muirhead RJ (2007). On a multiple endpoints testing problem. *Journal of Statistical Planning and Inference* 137:3416-3429.

Hung HM, Wang SJ (2009). Some controversial multiple testing problems in regulatory applications. *J Biopharm Stat.* 19:1-11.

Julious SA, McIntyre NE (2012). Sample sizes for trials involving multiple correlated must-win comparisons. Pharmaceut. Statist. Published online in Wiley Online Library.

Kong L, Kohberger RC, Koch GG (2004) Type I error and power in noninferiority/equivalence trials with correlated multiple endpoints: an example from vaccine development trials. *J Biopharm Stat.* 14: 893-907.

Kordzakhia, G, Siddiqui, O and Huque, M.F (2010), Method of balanced adjustment in testing co-primary endpoints. Statist. Med. 29: 2055–2066.

Offen W, et al. (2007). Multiple co-primary endpoints: medical and statistical solutions. *Drug Information Journal* 41:31-46.

Park, CG, Park T, Shin, DW (1996). A simple method for generating correlated binary variates, The American Statistician, 50: 306-310.

Rosenberger, WF, Lachin, JM (2002). Randomization in Clinical Trials: Theory and Practice. Wiley, New York.

Sankoh, AJ, D'Agostino, RB and Huque, MF (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Statist. Med. 22: 3133–3150.

Song, JX (2009). Sample size for simultaneous testing of rate differences in non-inferiority trials with multiple endpoints. Comp. Stat. & Data Anal. 53: 1201-1207.

Sozu T, Sugimoto T and Hamasaki T (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. Statist. Med. 29: 2169-2179.

Sozu T, Sugimoto T, Hamasaki T (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. J Biopharm Stat. 21:650-668.

Xiong C, Yu K, Gao F, Yan Y, Zhang Z (2005). Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. *Clinical Trials* 2:387-393.

Wang B, Ting N (2016). Sample size determination with familywise control of both type I and type II errors in clinical trials, J Biopharm Stat, 26:5, 951-965.

Wicklin, R (2013). Simulating Data with SAS. Cary, NC: SAS Institute Inc.

Zhang, Z (2012). Power and Sample Size Estimation in Clinical Trials with Multiple Co-primary Endpoints, In *JSM Proceedings*, Biopharmaceutical Section. Alexandria, VA: American Statistical Association. 542-553.

Zhang, Z (2017). Sample Size and Assurance Probability Calculation in Multi-regional Clinical Trials (draft).

Appendix 1

Power and sample size estimation by simulating test statistic vectors: a SAS program for $K = 2$ co-primary binary endpoints

```
* Input design parameters for power and sample size estimation;
%let pi_A1=0.30;  * Arm A response rate for endpoint 1;
%let pi_A2=0.25;  * Arm A response rate for endpoint 2;
%let rho_A12=0.7;* Arm A endpoints 1 and 2 correlation coefficient;
%let pi  C1=0.10;  * Arm C response rate for endpoint 1;
%let pi_C2=0.08;  * Arm C response rate for endpoint 2;
%let rho_C12=0.3; * Arm C endpoints 1 and 2 correlation coefficient;
%let q_A=0.6667;  * Proportion of subjects randomized to arm A;
%let q_C=0.3333;  * Proportion of subjects randomized to arm C;
%let nmin=230;     * Total sample size low bound;
%let nmax=250;    * Total sample size upper bound;
%let alpha=0.025;  * One-sided significant level;
%let ntrial=50000; * Total number of simualted trials;
%let seed=778899; * Seed for generating random numbers;

data d0;
  q_A = &q_A; q_C = &q_C; alpha=&alpha;
  z_alpha=-probit(alpha); ntrial=&ntrial;
  pi_A1 = &pi_A1; pi_A2 = &pi_A2; rho_A12 = &rho_A12;
  pi  C1 = &pi  C1; pi  C2 = &pi  C2; rho  C12 = &rho  C12;
  pi_AC1 = q_A*pi_A1 + q_C*pi_C1; pi_AC2 = q_A*pi_A2 + q_C*pi_C2;
  ph_AC1_0 = sqrt(pi_AC1*(1-pi_AC1)*(1/q_A + 1/q_C));
  ph_AC2_0 = sqrt(pi_AC2*(1-pi_AC2)*(1/q_A + 1/q_C));
  ph_AC1_1 = sqrt(pi_A1*(1-pi_A1)/q_A + pi_C1*(1-pi_C1)/q_C);
  ph_AC2_1 = sqrt(pi_A2*(1-pi_A2)/q_A + pi_C2*(1-pi_C2)/q_C);
```

```
  r_AC12 = (rho_A12*sqrt(pi_A1*(1-pi_A1)*pi_A2*(1-pi_A2))/q_A +
rho_C12*sqrt(pi_C1*(1-pi_C1)*pi_C2*(1-pi_C2))/q_C)/(ph_AC1_1*ph_AC2_1);
  fAC_1 = ph_AC1_0/ph_AC1_1; fAC_2 = ph_AC2_0/ph_AC2_1;
  gAC_1 = abs(pi_A1 - pi_C1)/ph_AC1_1;
  gAC_2 = abs(pi_A2 - pi_C2)/ph_AC2_1;
  CALL SYMPUT('r_AC12', PUT(r_AC12, best.));
run;

proc iml;
  Mean = {0, 0};
  I={1 0, 0 1};
  J={0 1, 1 0};
  Cov = I + &r_AC12*J;
  call eigen(egvalue, egvector, cov); print cov egvalue egvector;
  ntrial = &ntrial; call randseed(&seed);
  W = RandNormal(ntrial, Mean, Cov); varNames = "W1":"W2";
  create sd0 from W[colname=varNames]; append from W; close sd0;
quit;

data sd1; if _n_=1 then set d0; set sd0; run;

data sd2;
  set sd1;
  do n=&nmin to &nmax;
          wac1_alpha = fAC_1*z_alpha - gAC_1*sqrt(n);
          wac2_alpha = fAC_2*z_alpha - gAC_2*sqrt(n);
    output;
  end;
run;

proc sort data=sd2 out=sd3; by n ntrial; run;

data sd4;
  set sd3; by n; retain sigcnt;
  if first.n then sigcnt=0;
  sigcnt = sigcnt + (W1>=wac1_alpha)*(W2>=wac2_alpha);
  if last.n then do; sigpcent = sigcnt*100/ntrial; output; end;
  label n="Total Sample Size for both arms"
          sigpcent="Power (percent of significant simulated trials)";
run;

proc print; run;
```