東海大學統計研究所
碩士論文

二期臨床試驗之二階段分層隨機設計：條件方法

Two-stage Designs for Stratified Randomized Phase II Trials ：Conditional Approach

指導教授:張玉媚 博士

研究生:戴弘儀

中華民國一零七年十二月

致謝

本論文之完成，衷心感謝我的指導教授－張玉媚老師。在這研究所的一年半中老師總是細心的提點我，既使我對於觀念的理解較為緩慢，但是老師依然很有耐心地給予鼓勵並且一步一步的帶我去了解。也感謝在老師忙於授課、研究和家庭之餘，仍然撥了許多時間給予我指導以及幫助，帶領我完成論文以及口試。同時，我要感謝我的論文口試委員沈葆聖老師以及陳春樹老師，感謝兩位老師在百忙之中抽空前來口試指導，提供我寶貴的建議與協助，使得論文更加完善。

最後我要感謝所有東海統計研究所的老師、助教以及研究室的同學們，很幸運能有你們在我就讀研究所的這段時間裡陪伴我學習並且給予幫助，還有我的家人，一直以來的鼓勵與支持，讓我能順利的取得學位。

戴弘儀

謹致於東海大學統計研究所

中華民國一零七年十二月

Contents

Table Listing

Figure Listing

ABSTRACT

Two-stage designs have been widely used in phase II clinical trials to evaluate the efficacy and safety of the study treatment. A common primary endpoint is a binary (yes/no) patient response to treatment. In some cases, the patient response distribution for a phase II clinical trial is heterogeneous, making it desirable to stratify patients into subgroups according to different prognostic factors. In this article, for a two-arm stratified randomized phase II clinical trial, we consider two-stage designs and propose three testing procedures to compare the response rates between two treatments. The first procedure is based on the weighted average of the stratum-specific differences between treatment response rates. The second and third procedures are based on the estimated relative risk and odds ratio, respectively, under the assumption of a common odds ratio over the strata. We consider conditional approach and present a simulation-based algorithm by modifying the algorithm in London and Chang (2005) to determine the parameters in designs to achieve the desired power at the nominal level. Simulation results show that the split-levels of type I and type II errors and randomization ratio have a crucial impact on the overall sample size required. Decreasing the split-level or increasing the randomization ratio at the first-stage can result in a smaller total sample size if early termination after the first-stage does not occur. In terms of the total sample size required, the INVAR-weighted test outperforms the other tests when the odds ratio or the true difference between two response rates is constant across strata. When neither odd ratio nor the difference between two response rates is constant across the strata, the INVAR-weighted test also performs well when the randomization ratio is large.

KEYWORDS: Odds ratio; Response rate; Sample size; Stratification; Two-stage design

1. Introduction

One of the objectives of a phase II clinical trial is to evaluate the effect of an experimental treatment and decide whether it is promising to be studied in a larger-scale phase III trial. Phase II clinical trials are often single-arm studies and the endpoint is typically a binary patient response such that the objective response rate can be used to assess the effect of an experimental treatment. The patient population or a phase II clinical trials can be heterogeneous across subgroups. Since the response rates differ across the strata, it is inappropriate to conduct the binomial test under the assumption that the number of responses follows a binomial distribution with the same response probability for all patients. On the other hand, sample size in a phase II clinical trial is usually small such that it is inefficient to conduct independent binomial tests within subgroups. In this situation, it will be desirable to stratify patients into subgroups according to different prognostic factors, such as age, gender, disease stage, and/or other risk factors, which are expected to have quite different response rates. For single-arm stratified phase II trials, London and Chang (2005) have proposed conditional and unconditional approaches for generating sample sizes and stopping boundaries that provide one-stage and two-stage designs with the desired power at nominal level. Their proposed test statistic was based on the difference of the observed total number of responses over strata and the corresponding expected number of responses under the null hypothesis. Since the unconditional approach requires the information on the proportions of patients for all strata, they suggested using the conditional approach, where initial estimates of the proportions of patients are needed to compute the sample size before the study. One advantage of the conditional approach is that decision boundaries can be changed according to the observed numbers of patients in the defined strata. They also proposed a simulation-based method to determine the parameters in designs. Chang et al. (2012) pointed out that the test statistic proposed by London and Chang (2005) is equivalent to an equal-weight (common odds ratios) linear combination of the numbers of respondents in the strata, leading to some loss of power. To improve the power of test, they proposed an unequal-weight test statistic based on Neyman-Pearson lemma. Their numerical results indicate that the proposed test is more powerful than London and Chang's test (2005). Other studies of the single-arm phase II clinical trial with stratification can be seen in Thall et al. (2003), Chang et al. (2011) and Jung et al. (2012).

The single-arm phase II trial designs for evaluating each experimental treatment individually are limited by outcome-trial effect confounding arising from the incapability of separating trial effects (such as patient selection, trial eligibility, and treatment locations) from treatment effect on clinical outcomes. Instead, randomized designs to experimental regimens, using a control arm when necessary, offer an attractive proposition by ensuring better patient comparability and reducing confounding between outcome and trial effects. For more than two decades, there has been interest in utilizing phase II trials with randomization against a standard-treatment control arm to provide greater assurance than afforded by comparison to historic controls that the new regimen is promising and warrants further evaluation (Rubinstein et al. 2005). Simon et al. (1985) described the randomized Phase II trials with a control arm. Jung (2008) and Thall et al. (1989) proposed different two-stage designs for randomized phase II trials with a control treatment. In this article, for a two-arm stratified randomized phase II clinical trial, we consider two-stage designs and propose three testing procedures to compare the response rates between two treatments. The first procedure is based on the weighted average of the stratum-specific differences between treatment response rates. The second and third procedures are based on the estimated relative risk and odds ratio, respectively, under the assumption of a common odds ratio over the strata. Since in practice an accurate estimate of the proportions of patients for strata is usually not available, we consider conditional approach and present a simulation-based algorithm by modifying the algorithm in London and Chang (2005) to determine the parameters in designs to achieve the desired power at the nominal level. Thus, the conditional approach based on the first procedure is an extension of the conditional method for a single-arm trial in London and Chang (2005).

The rest of the article is organized as follows. In Section 2, we review three test statistics for comparing two binomial proportions from stratified samples. In Section 3, we consider conditional approach in two-stage designs for a two-arm stratified randomized phase II clinical trial. We present a simulation-based algorithm to find the parameters in the design to achieve the desired power at the nominal level. Some numerical examples under various settings of expected response rates in the experimental and the control treatment groups for the proposed design are presented in Section 4. Finally, conclusions and discussions are given in Section 5.

2. The testing procedures

Suppose that patients can be stratified into q strata. Let N_i^e be the number of patients and X_i^e be the numbers of responses in the *i*th stratum of the experimental treatment group. Let N_i^c be the number of patients and X_i^c be the numbers of responses in the *i*th stratum of the control group. Conditional on the observed numbers of patients $N_i^e = n_i^e$ and $N_i^c = n_i^c$, $i = 1, \dots, q$, we assume that $X_1^e, \dots, X_q^e$ and $X_1^c, \dots, X_q^c$ are independent binomial random variables with

$$X_i^e \sim Binomial(n_i^e, \pi_i^e) \quad \text{and} \quad X_i^c \sim Binomial(n_i^c, \pi_i^c),$$

where π_i^e and π_i^c are the expected response rates of the experimental and the control treatments in the ith stratum, respectively.

2.1. Test based on the difference between response rates

Let $\eta_i = \pi_i^e - \pi_i^c$ be the true difference between the experimental and the control response rates in the ith stratum, $i = 1, \ldots, q$. The true overall treatment effect is given by $= \Sigma_{i=1}^q P_i \eta_i$, where P_i is the true proportion of patients from the ith stratum if the entire target population had been enrolled ($\Sigma_{i=1}^q P_i = 1$). The problem of testing the hypothesis that experimental treatment has larger response rate than the control treatment can be unified as the following hypothesis:

$$H_0^1: \pi_i^e = \pi_i^c = \pi_i^0 \ \text{ vs } \ H_1^1: \pi_i^c = \pi_i^0, \pi_i^e > \pi_i^c, \text{ at least one } i, i=1, \ldots, q,$$

stratum-specific with the desired significant level α, and power (1- β) evaluated at $\pi_i^e = \pi_i^c + \Delta_i$, where Δ_i is the specified improvement in response rate in strum i we want to detect.

The nature estimate of η is given by

$$\hat{\eta} = \Sigma_{i=1}^q w_i \hat{\eta}_i = \Sigma_{i=1}^q w_i \left(\frac{X_i^e}{n_i^e} - \frac{X_i^c}{n_i^c} \right),$$

where w_i is the weight assigned to the ith stratum satisfying $\Sigma_{i=1}^q w_i = 1$. There are two common methods to determine. One method is the harmonic means of the samples size (SSIZE), that is

$$w_i = \frac{(n_i^e n_i^c)/(n_i^e + n_i^c)}{\Sigma_{i=1}^q (n_i^e n_i^c)/(n_i^e + n_i^c)},$$

which are also referred to as the Cochran-Mantel-Haenszel (Cochran 1954; Mantel-Haenszel 1959) weights for comparing two independent proportions with stratification. The other one is the reciprocals of the variances of the stratum-specific differences (INVAR), that is

$$w_i = \frac{V_i^{-1}}{\Sigma_{i=1}^{q} V_i^{-1}} \ ,$$

where

$$V_i = \frac{(X_i^e/n_i^e)/(1 - X_i^e/n_i^e)}{n_i^e} + \frac{(X_i^c/n_i^c)/(1 - X_i^c/n_i^c)}{n_i^c}$$

is the variance of $\hat{\eta}_i$. For the SSIZE weighting method, a larger weight is assigned to strata with a large number of patients compared to that with a small number of patients. For the INVAR weighting method, a larger weight is assigned to strata with a small value of the estimated variance of the difference between the response rates compared to that with a large value. The estimate $\hat{\eta}$ based on the SSIZE weighting method is generally unbiased or approximately unbiased. Although the estimate $\hat{\eta}$ based on the INVAR weighting method is usually a biased estimate of η when η_i's are not constant, it has minimum variance (Mehrotra and Railkar 2000). Radhakrishna (1965) showed that the SSIZE weighting method is optimal if the odds ratio $\pi_i^e(1 - \pi_i^c)/\{\pi_i^c(1{-}\pi_i^e)\}, i = 1, \dots, q$, are constant, and the INVAR weighting method is optimal if $\eta_i, i = 1, \dots, q$, are constant.

Because the wrong choice of weighting method may lead to the loss in efficiency, Mehrotra and Railkar (2000) proposed the minimum risk (MR) weighting method by minimizing the average squared error loss of $\hat{\eta}$, $E(\hat{\eta} - \eta)^2$, that is

$$w_i = \frac{b_i}{\Sigma_{i=1}^{q} V_i^{-1}} - \left(\frac{a_i V_i^{-1}}{\Sigma_{i=1}^{q} V_i^{-1} + \Sigma_{i=1}^{q} a_i \eta_i V_i^{-1}}\right)\left(\frac{\Sigma_{i=1}^{q} b_i \eta_i}{\Sigma_{i=1}^{q} V_i^{-1}}\right),$$

where $a_i = \eta_i \Sigma_{i=1}^{q} V_i^{-1} - \Sigma_{i=1}^{q} \eta_i V_i^{-1}$, and $b_i = V_i^{-1}(1 + a_i \Sigma_{i=1}^{q} P_i \eta_i)$, which reduces to the INVAR weights when η_i, $i = 1, \dots, q$, are constant across strata. The estimate $\hat{\eta}$ based on the MR weights is more precise and less biased relative to the SSIZE and the INVAR weighting methods.

For testing $H_0^1: \pi_i^e = \pi_i^c = \pi_i^0, i = 1, \dots, q$, the test statistic denoted by T_{Diff} is given by

$$\mathrm{T}_{Diff} = \frac{\hat{\eta}}{\sqrt{\Sigma_{i=1}^{q} \mathrm{w}_i^2 \widehat{var}(\hat{\eta}_i)}} \tag{1}$$

where

$$\widehat{Var}(\hat{\eta}_i) = \left(\frac{1}{n_i^e} + \frac{1}{n_i^c}\right)\left(\frac{X_i^e + X_i^c}{n_i^e + n_i^c}\right)\left(1 - \left(\frac{X_i^e + X_i^c}{n_i^e + n_i^c}\right)\right),$$

and the test statistic converges to the standard normal distribution as the sample size for each treatment in each stratum tends to infinity.

2.2. Test based on the odds ratio

An alternative testing procedure is based on the odds ratio, which is used in the analysis of stratified two-by-two tables. The odd ratio in the ith stratum is given by

$$\vartheta_i = \frac{\pi_i^e (1 - \pi_i^c)}{\pi_i^c (1 - \pi_i^e)}$$

Let $\theta_i = \log\vartheta_i$ be natural logarithm of ϑ_i. Testing hypothesis H_0^1 versus H_1^1 is equivalent to the following hypothesis:

$$H_0^2: \theta_1 = \cdots = \theta_q = 0 \text{ vs } H_1^2: \theta_i > 0\text{, at least one } i, i=1,\dots, q,$$

with the desired significant level α, and power $(1 - \beta)$ evaluated at $\theta_i = \delta_i > 0$,where δ_i is the specified improvement in logarithm of odds ratio in stratum i we want to detect.

Assuming that the odds ratio is constant, i.e., $\theta_1 = \cdots = \theta_q = \theta$, the estimate of $\hat{\theta}$ is given by

$$\hat{\theta} = \log\left(\frac{\Sigma_{i=1}^{q} X_i^e (n_i^c - X_i^c)/(n_i^e + n_i^c)}{\Sigma_{i=1}^{q} X_i^c (n_i^e - X_i^e)/(n_i^e + n_i^c)}\right),$$

and the variance estimate of $\hat{\theta}$ is given by

$$\widehat{var}(\hat{\theta}) = \frac{\Sigma_{i=1}^{q} S_i R_i}{2(\Sigma_{i=1}^{q} R_i)^2} + \frac{\Sigma_{i=1}^{q}(S_i U_i + Q_i R_i)}{2(\Sigma_{i=1}^{q} R_i)(\Sigma_{i=1}^{q} U_i)} + \frac{\Sigma_{i=1}^{q} Q_i R_i}{2(\Sigma_{i=1}^{q} U_i)^2},$$

where $S_i = (X_i^e + n_i^c - X_i^c)/(n_i^e + n_i^c)$, $Q_i = (X_i^c + n_i^e - X_i^e)/(n_i^e + n_i^c)$, $R_i = X_i^e(n_i^c - X_i^c)/(n_i^e + n_i^c)$ and $U_i = X_i^c(n_i^e - X_i^e)/(n_i^e + n_i^c)$ (Jennison and Turnbull 1991).

When H_0^2 is true, the following test statistic

$$T_{OR} = \frac{\hat{\theta}}{\sqrt{\widehat{var}(\hat{\theta})}} \tag{2}$$

converges to the standard normal distribution as the sample size for each treatment in each stratum tends to infinity (Jennison and Turnbull 1991). Given π_i^0 and δ_i, the previous hypothesis $H_0^2: \theta_1 = \cdots = \theta_q = 0$ vs $H_1^2: \theta_i = \delta_i > 0$ is equivalent to the hypothesis $H_0^1: \pi_i^e = \pi_i^c = \pi_i^0$ vs $H_1^1: \pi_i^c = \pi_i^0, \pi_i^e + \Delta_i, i = 1, \dots, q$, with

$$\Delta_i = \pi_i^0(1 - \pi_i^0)(\exp(\delta_i) - 1)/(1 + \pi_i^0(\exp(\delta_i) - 1)).$$

2.3. Test based on relative risk

The other alternative testing procedure is based on the relative risk, which is also commonly used in the binary response as a measure of endpoints. The relative risk in the *i*th stratum is given by $\varphi_i = \pi_i^e/\pi_i^c$. Hence, the true overall treatment effect is given by $\varphi = \sum_{i=1}^{q} P_i \varphi_i$. Testing hypothesis H_0^1 versus H_1^1 is equivalent to the following hypothesis:

$$H_0^3: \varphi_1 = \cdots = \varphi_q = 1 \text{ vs } H_1^3: \varphi_i > 1, \text{at least one } i, i=1,\dots,q,$$

with the desired significant level α, and power (1- β) evaluated at $\varphi_i = \emptyset_i > 1$, where $\emptyset_i$ is the specified improvement in relative risk in stratum i we want to detect. We use the Mantel-Haenszel type risk ratio (Rothman and Boice 1979; Tarone 1981) to estimate the overall relative risk across strata, which is given by

$$\hat{\varphi} = \frac{\sum_{i=1}^{q} X_i^e n_i^c / (n_i^e + n_i^c)}{\sum_{i=1}^{q} X_i^c n_i^e / (n_i^e + n_i^c)},$$

and the asymptotic variance of $\hat{\varphi}$ can be estimated by

$$\widehat{Var}(\hat{\varphi}) = \frac{\sum_{i=1}^{q} X_i^e \left(\frac{n_i^c}{n_i^e + n_i^c}\right)^2 + \hat{\varphi}^2 \sum_{i=1}^{q} X_i^c \left(\frac{n_i^e}{n_i^e + n_i^c}\right)^2}{\left(\sum_{i=1}^{q} X_i^c n_i^e / (n_i^e + n_i^c)\right)^2}.$$

When H_0^3 is true, the following test statistic T_{RR}

$$T_{RR} = \frac{\hat{\varphi} - 1}{\sqrt{\widehat{Var}(\hat{\varphi})}} \quad (3)$$

converges to the standard normal distribution as the sample size for each treatment in each stratum tends to infinity. Given π_i^0 and φ_i, the previous hypothesis $H_0^3 \colon \varphi_1 = \cdots = \varphi_q = 1$ vs $H_1^3 \colon \varphi_i > 1, i = 1, \ldots, q$ is equivalent to the hypothesis $H_0^1 \colon \pi_i^e = \pi_i^c = \pi_i^0$ vs $H_1^1 \colon \pi_i^c = \pi_i^0, \pi_i^e + \Delta_i, i = 1, \ldots, q$, with

$$\Delta_i = \pi_i^0 (\varphi_i - 1).$$

3. The two-stage designs: conditional approach

Since clinician usually do not have an accurate estimates of $P_1, \ldots P_q$, one-stage design and the unconditional approach in two-stage design may be impractical. Thus, we consider two-stage designs and conditional approach. Suppose that patients are stratified into q strata for a two-arm stratified randomized phase II clinical trial. First, we briefly describe the conditional approach. Assume that the initial rough estimates of $P_1, \ldots P_q$ are available. For testing H_0 versus H_1, based on the desired type I error and power, M_1^e and M_1^c patients are randomly assigned to receive the experimental and the control treatments at the first stage. At the second stage, based on the observed accrual rate for each stratum, the additional sample sizes for the experimental and control treatments are M_2^e and M_2^c, respectively.

Let n_{ij}^c and n_{ij}^e be the number of patients in the ith stratum at the jth stage of the control group and the treatment group, respectively. Also let X_{ij}^c and X_{ij}^e be the number of responses among the n_{ij}^c and n_{ij}^e patients, respectively. After $M_1^e + M_1^c$ patients have entered the study at the first stage, the test statistic, denoted by T_1, which can be one of the test statistics proposed in Section 2.1-2.3, and will be calculated based on the response data of the $M_1^e + M_1^c$ patients, where the observed numbers of patients $n_{11}^e, \dots, n_{q1}^e$ and $n_{11}^c, \dots, n_{q1}^c$, for treatment and control group, respectively. If $T_1 < a_1$, then we fail to reject the null hypothesis, declare the experimental treatment is not promising and the study is stopped; if $T_1 > b_1$, then we reject the null hypothesis and the study is also stopped; if $a_1 \leq T_1 \leq b_1$, then the accrual will be continued for the second stage, where a_1 is the largest real number satisfying

$$P_{H_1}(T_1 < a_1 | n_{i1}^e, n_{i1}^c, i = 1, \dots, q) \approx \tau_2 \beta, \tag{4}$$

and b_1 is the smallest real number satisfying

$$P_{H_0}(T_1 > b_1 | n_{i1}^e, n_{i1}^c, i = 1, \dots, q) \approx \tau_1 \alpha, \tag{5}$$

where τ_1 and τ_2 can be chosen based on the guidance of Fleming et al. (1982) and Chang et al. (1998).

If the accrual continues to the second stage for the next $(M_2^e + M_2^c)$ patients, then the test statistic, denoted by T_2, which is the same test statistics as stage 1, will be calculated. If $T_2 \leq b_2$, then we fail to reject the null hypothesis and conclude that the experimental treatment is not promising; if $T_2 > b_2$, then we reject the null hypothesis and claim that the experimental treatment is promising, where b_2 is the smallest real number satisfying

$$P_{H_0}(T_1 > b_1 | n_{i1}^e, n_{i1}^c, i = 1, \ldots, q)$$

$$+P_{H_0}\left(a_1 \le T_1 \le b_1, T_2 > b_2 \middle| n_{ij}^e, n_{ij}^c, i = 1, \ldots, q, j = 1,2\right) \le \alpha. \qquad (6)$$

The power of the test is

$$\text{Power} = P_{H_1}(T_1 > b_1 | n_{i1}^e, n_{i1}^c, i = 1, \ldots, q)$$

$$+P_{H_1}\left(a_1 \le T_1 \le b_1, T_2 > b_2 \middle| n_{ij}^e, n_{ij}^c, i = 1, \ldots, q, j = 1,2\right). \qquad (7)$$

The design with the decision boundaries a_1, b_1 and b_2 guarantees that significant level does not exceed α.

The design parameters need to be determined before the study begins. We propose a simulation-based method to determine sample sizes M_1^c, M_1^e, M_2^c, M_2^e and the boundaries a_1, b_1, and b_2 for achieving the desired power at the nominal level, according to the context of the hypothesis and the rough estimates of $P_1, \ldots P_q$. Let r_1 and r_2 denote the randomization ratios for the first stage and the second stage, respectively, i.e., $M_1^e/M_1^c \approx r_1$, $M_2^e/M_2^c \approx r_2$ and $M_2^c \approx k \times M_1^c$. This design is referred to as an unbalanced design if $r_1 \ne 1$ or $r_2 \ne 1$ and more patients will be assigned to the experimental treatment group if $r_1 > 1$ or $r_2 > 1$. Using the approach in London and Chang (2005), we propose the following simulation-based algorithm to determine the design parameters:

1. Set the initial values of M_1^c to be smaller than the anticipated sample size, and set M_2^c to be the integral part of $k \times M_1^c$. Let M_1^e and M_2^e be the integral parts of $r_1 \times M_1^c$ and $r_2 \times M_2^c$, respectively.

2. For $j = 1, 2$, $i = 1, \ldots, q$-1, and x=e, c, let n_{ij}^x be the nearest integral of $M_j^x \times P_i$, and $n_{qj}^x = M_j^x - \sum_{i=1}^{q-1} n_{ij}^x$.

3. Generate the binomial random variable X_{ij}^{e} with sample size n_{ij}^{e} and response rate π_i^e, and generate the binomial random variable X_{ij}^{c} with sample size n_{ij}^{c} and response rate π_i^c, where π_i^e and π_i^c are defined in the null hypothesis.

4. Compute the test statistics T_1 and T_2 based on the values n_{ij}^{e}, n_{ij}^{c}, x_{ij}^{e}, and x_{ij}^{c} obtained in steps 2 and 3.

5. Repeat Step 3 and Step 4, say 50,000 times, we can obtain the estimate of the joint distribution of (T_1, T_2) and the estimate of the marginal distribution of T_1 under the null hypothesis, the latter can be used to obtain b_1 according to (5).

6. Repeat Step 3 and Step 4, say 50,000 times, where π_i^e and π_i^c are defined in the alternative hypothesis. Then we can obtain the estimates of the joint distribution of (T_1, T_2) and the marginal distribution of T_1 under the alternative hypothesis, and the latter can be used to obtain a_1 according to (4).

7. After obtaining a_1 and b_1, the estimate of the joint distribution of (T_1, T_2) under the null hypothesis can be used to obtain b_2 according to (6).

8. Given a_1, b_1, b_2 obtained in the previous steps, we use the estimate of the joint distribution of (T_1, T_2) under the alternative hypothesis to evaluate the desired power (7). If the test power is lower than the desired power, then M_1^c is set to be $M_1^c + 1$.

9. Repeat Steps 2-8 until the desired power requirement is satisfied.

Jung (2008) pointed out that the unbalanced two-stage design usually requires larger total sample size compared with the balanced design. The discussion on the ratio of the experimental treatment sample size and the control treatment sample size can be seen in (Wittes 2002). In next section, some numerical examples are given to demonstrate how the total sample size is impacted by various settings of k and r_i.

Now, let the $\hat{P}_i$ be the estimate of P_i based on n_{ij}^{e}, n_{ij}^{c} as follows

$$\hat{P}_i = \frac{\sum_{j=1}^{2}\left(n_{ij}^{e} + n_{ij}^{c}\right)}{\sum_{j=1}^{2}\sum_{i=1}^{q}\left(n_{ij}^{e} + n_{ij}^{c}\right)}.$$

Notice that the determination of sample sizes M_1^e, M_1^c, M_2^e, and M_2^c depends on the initial estimates of the P_i's. If $\hat{P}_i$'s do not differ from P_i's too much, the actual power will be close to the desired level. Instead, if $\hat{P}_i$ differs from P_i a lot for some i, the desired power may not be achieved so that an adjustment of the sample size is required. Following the advice of London and Chang (2005), we could enroll more patients into the experimental/control group at the second stage for achieving the desired power.

4. Numerical examples

We compared the sample sizes based on the test statistic T_{Diff} (1) with three different weighting methods. At the same time, we also presented the sample sizes based on the test statistic T_{OR} (2) and T_{RR} (3). We considered three strata with equal proportion $P_1 = P_2 = P_3 = 1/3$. Under the balanced design with the same number of patients assigned into each treatment group, i.e. k =1 and $r_1 = r_2 = 1$, we considered three different scenarios in the numerical study, where the expected response rates of the control treatment $(\pi_1^c, \pi_2^c, \pi_3^c) = (\pi_1^0, \pi_2^0, \pi_3^0)$ were set as (0.4, 0.2, 0.1), and (0.6, 0.3, 0.1).

Scenario 1: Equal difference. In this scenario, the true difference between the experimental response rate and the control response rate is set as $\Delta_1 = \Delta_2 = \Delta_3 = \Delta$=0.20, 0.23 and 0.25.

Scenario 2: Equal odds ratio. In this scenario, the logarithm of odds ratio is set as $\delta_1 = \delta_2 = \delta_3 = \delta$ =1.1, 1.25 and 1.5.

Scenario 3: Unequal difference. In this scenario, the true difference between the experimental response rate and the control response rate is not constant across strata, implying different odds ratio between the strata.

Furthermore, we consider the forth scenario with an unbalanced design under various settings for k=(0.5, 1.0, 2.0) and (r_1, r_2)=(1, 1), (1, 2), (2, 1) and (2, 2), in which the expected response rates of the control treatment are set as (0.4, 0.2, 0.1) and the true difference between the experimental and the control response rates are set as (0.35, 0.24, 0.12), which is not constant across strata. For each scenario, Type I error probability α and Type II error probability β were set to be 0.05 and 0.2, respectively. To decide the critical points a_1, b_1, the split-level for both errors is set to be 50% at the first stage, i.e., $\tau_1 = \tau_2$ =0.5. The total sample size N, M_1^c, M_1^e, M_2^c, M_2^e, a_1, b_1, b_2, and the desired power are presented in Tables 1-4. Notice that the total sample size N refers to the minimum sample size (MS) required for achieving the desired α and β if early termination after the first stage does not occur. Since the impact of r_1 and r_2 on total sample size is shown in Figure 1, Table 4 only lists the results for $r_1 = r_2 = 1$.

Table 1 indicates that to achieve the desired level the required sample size based on the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the smallest when the true difference between the experimental and the control response rates is constant across strata. Due to violation of the assumption of the constant odd ratios, the SSIZE-weighted T_{Diff} and T_{OR} require larger sample size to achieve the desired level, in particular, the T_{RR} requires the largest sample size to achieve the desired level. Similarly, the required sample sizes get smeller, as the value of Δ_i increases.

From Table 2, we can observe that to achieve a desired level the required sample size of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} is the smallest when the odds ratio is constant across strata. When the values of δ increases, and the required sample sizes of the INVAR-weighted T_{Diff}, MR-weighted T_{Diff}, SSIZE-weighted T_{Diff} and T_{OR} are getting closer and the required sample size based on the test statistic T_{RR} is the largest.

When neither constant odds ratio nor constant difference holds, Table 3 shows that to achieve the desired level when the difference between Δ_1, Δ_2 and Δ_3 (or δ_1, δ_2 and δ_3) is large, the required sample sizes of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the smallest and the test statistic T_{RR} is the largest, respectively. However, if the difference between Δ_1, Δ_2 and Δ_3 (or δ_1, δ_2 and δ_3) is small, the required sample size based on the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are smaller than or equal to that based on the SSIZE -weighted T_{Diff} and the test statistic T_{OR}.

The results for the unbalanced design in Table 4 and Figure 1 indicate that to achieve the desired level, the required sample sizes of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the smallest and on the test statistic T_{RR} is the largest. The required sample size based on the SSIZE -weighted T_{Diff} is close to the test statistic T_{OR}. The required sample sizes based on all the test statistic T is the smallest when k=0.5 and (r_1, r_2)=(2, 1).

Moreover, we also consider the scenarios with τ_1 and τ_2 set to 25% and 25% at the first stage and various setting of (r_1, r_2). We consider four scenarios: (Scenario 1a): Equal difference (Scenario 2a): Equal odds ratio (Scenario 3a): Unequal difference under with k =1, and (Scenario 4a): unbalanced designs under various settings with k=(0.5, 1.0, 2.0) and (r_1, r_2)=(0.5, 0.5), (0.5, 1), (1, 1), (1, 2), (2, 1) and (2, 2) under $(\pi_1^0, \pi_2^0, \pi_3^0)$= (0.4, 0.2, 0.1) and ($\Delta_1$, Δ_2, Δ_3) = (0.35, 0.24, 0.12). The results are shown in Tables 5-8. Since the impact of r_1 and r_2 on total sample size is shown in Figure 2, Table 8 only lists the results for (r_1, r_2)= (2, 1).

From Table 5, to achieve the desired level the required sample size based on the

INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the smallest and the T_{RR} require the largest sample size. The required sample size based on the SSIZE-weighted T_{Diff} is close to the test statistic T_{OR} and require larger sample size to achieve the desired level. When (r_1, r_2)=(2, 1) the required sample size is smaller than that under (r_1, r_2)= (1, 1). When (τ_1, τ_2)=(0.25, 0.25), the required total sample size is smaller than that under (τ_1, τ_2)=(0.5, 0.5). Similarly, the required sample sizes get smaller, as the value of Δ_i increases.

From Table 6, we can observe that to achieve a desired level the required sample size of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} is the smallest and the test statistic T_{RR} is the largest. For $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.4, 0.2, 0.1), when the values of δ increases, the required sample sizes of the SSIZE-weighted T_{Diff} , the test statistic T_{OR} and T_{RR} are the same. When the $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.4, 0.2, 0.1) and the δ=1.1, the required sample size based on the SSIZE-weighted T_{Diff} under $r_1 = r_2 = 1$ are the same as that under $r_1 = 2,\ r_2 = 1$. When the $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.6, 0.3, 0.1) and the δ =1.1, the required sample size based on the INVAR-weighted T_{Diff} is the smallest and the test statistic T_{RR} is getting larger as (r_1, r_2)=(2, 1). In addition, when the (r_1, r_2)=(2, 1), the required sample size based on all the test statistics is smaller than that under (r_1, r_2)=(1, 1).

From Table 7, to achieve the desired level the required sample sizes of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the smallest and the test statistic T_{RR} is the largest. When the difference between Δ_1, Δ_2 and Δ_3 (or δ_1, δ_2 and δ_3) is small, the required sample size based on the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are smaller than or equal to that based on the SSIZE -weighted T_{Diff} and the test statistic T_{OR}. And to achieve the desired level when the (r_1, r_2)=(2, 1), the required sample size is the smaller than when (r_1, r_2)=(1, 1) .When the difference between Δ_1,

Δ_2 and Δ_3 (or δ_1, δ_2 and δ_3) increases, the required sample size based on the SSIZE-weighted T_{Diff} and T_{OR} under the $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.6, 0.3, 0.1) but do not vary a lot as randomization ratio r_1 varies.

The results for the unbalanced design in Table 8 and Figure 2 indicate that to achieve the desired level, the required sample sizes of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the same and the smallest. The required sample sizes based on the test statistic T_{RR} is the largest. The required sample size based on the SSIZE -weighted T_{Diff} is close to the test statistic T_{OR}. When k=1 and (r_1, r_2)=(2, 2) the required sample sizes based on all the test statistics are the smallest.

5. Conclusions and Discussions

Under a two-stage design for stratified randomized two-arm phase II clinical trials, we have proposed three testing procedures to compare the response rates between two treatments. We have also developed a simulation-based algorithm to find the parameters in designs to achieve the desired power at the nominal level. Based on simulation results, we observe that to achieve the desired level, the required sample size of the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} is the smallest if the odds ratio is constant across strata, and the required sample size based on all the test statistics decrease as δ increases. The required sample size based on the INVAR- and MR-weighted T_{Diff} is the smallest if the true difference between two response rates is constant across strata, and the required sample size based on all the test statistics decrease as Δ increases. When the odd ratio and the difference between two response rates are not constant across the strata, the required sample sizes based on the INVAR-weighted T_{Diff} and MR-weighted T_{Diff} are the smallest and the test statistic T_{RR} is the largest, respectively in the case of large difference between Δ_1, Δ_2 and Δ_3 (or δ_1, δ_2 and δ_3). However, when the difference between Δ_1, Δ_2 and Δ_3 (or δ_1, δ_2 and δ_3) is small, the required sample size of the INVAR-weighted T_{Diff}, MR-weighted T_{Diff}, SSIZE-weighted T_{Diff} and T_{OR} are close. We also observe that the differences r_1 and r_2 based on the test statistics, the required sample sizes become smaller as the (r_1,r_2) increases. When the $\tau_1 = \tau_2 = 0.5$ and *k*=0.5, the required sample size based on all the test statistics is the smallest. When the $\tau_1 = \tau_2 = 0.25$ and *k*=1, the required sample size based on all the test statistics is the smallest.

The proposed conditional approach under a two-stage design can be extended to a multi-stage design for the stratified randomized two-arm trial. In a randomized phase II cancer clinical trial, sometimes the primary endpoint is the survival time, such as the progression-free survival time or the overall survival time (Sperduto et al. 2012). In this case, it may be worthwhile to develop a conditional approach for a two-arm stratified randomized phase II clinical trial.

References

Chang, M. N., Hwang. I. K., and Shih, W. J. (1998), "Group sequential designs using both type I and type II error probability spending functions", *Communications in Statistics-- Theory and Methods* 27, 1323-1339.

Chang, M. N., Jung, S. H., and Wu, S. S. (2011), "Two-stage designs with additional futility tests for phase II clinical trials with heterogeneous patient populations", *Sequential Analysis* 30, 338-349.

Chang, M. N., Shuster, J. J., and Hou, W. (2012), "Improved two-stage tests for stratified phase II cancer clinical trials", *Statistics in Medicine* 31, 1688-1698.

Cochran, W. G. (1954), "Some methods for strengthening the common chi-square tests", *Biometrics* 10, 417-451.

Jennison, C., and Turnbull, B. W. (1991), "A note on the asymptotic joint distribution of successive Mantel-Haenszel estimates of the odds ratio based on accumulating data", *Sequential Analysis* 10, 201-209.

Jung, S. H. (2008), "Randomized phase II trials with a prospective control", *Statistics in Medicine* 27, 568-583.

Jung, S. H., Chang, M. N., and Kang, S. J. (2012), "Phase II cancer clinical trials with heterogeneous patient populations" *Journal of Biopharmaceutical Statistics* 22, 312-328.

Fleming, T. R., (1982), "One-sample multiple testing procedures for phase II clinical trials", *Biometrics* 38, 143-151.

London, M. B., and Chang, M. N. (2005), "One- and two-stage designs for stratified phase II clinical trials" *Statistics in Medicine* 24, 2597-2611.

Mehrotra, D. V., and Raikar, R. (2000), "Minimum risk weights for comparing treatments in stratified binomial trials", *Statistics in Medicine* 19, 811-825.

Mantel, N., and Haenszel, W. (1959), “Statistical aspects of the analysis of data from retrospective studies of disease”, *Journal of the National Cancer Institute* 22, 719-748.

Radhakrishna, S. (1965), “Combination of results from several 2x2 contingency tables”, *Biometrics* 21, 86-98.

Rothman K. J, Boice J. D. (1979), Epidemiologic Analysis with a Programmable Calculator**.** Washington, DC: NIH Publication 21, 79-1649.

Rubinstein, L.V., Korn, E.L., Freidlin, B., Hunsberger, S., Ivy, S.P., and Smith, M.A. (2005), “Design issues of randomized phase II trials and a proposal for phase II screening trials”, *Journal of Clinical Oncology* 23, 7199-7206.

Simon, R., Wittes, R., and Ellenberg, S. (1985), “Randomized phase II clinical trials”, *Cancer Treatment Reports* 69, 1375-1381.

Sperduto P.W., Kased N., Roberge D., et al. (2012), “Effect of tumor subtype on survival and the graded prognostic assessment for patients with breast cancer and brain metastases”, *International Journal of Radiation Oncology Biology Physics* 82, 2111-2117.

Tarone, R. E. (1981), “On summary estimators of relative risk”, *Journal of Chronic Dieseaes* 34, 463-468.

Thall, P. F., Simon, R., and Ellenberg, S. (1989), “A two-stage design for choosing among several experimental treatments and a control in clinical trials”, *Biometrics* 45, 573-547.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, C. H., and Benjamin, R. S. (2003), “Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes”, *Statistics in Medicine* 22, 763-780.

Wittes, J. (2002), “Sample Size Calculations for Randomized Controlled Trials”, *Epidemiologic Reviews* 24, 39-53.

Table 1. Sample sizes under $\Delta_1 = \Delta_2 = \Delta_3 = \Delta$ under k=1, (τ_1, τ_2)=(0.5, 0.5) and (r_1, r_2)=(1, 1).

$(\pi_1^0,\pi_2^0,\pi_3^0)$	Δ	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.4,0.2,0.1)	0.2	Tss	138	36	36	33	33	0.568	1.884	1.670	0.801
		Tinv	114	30	30	27	27	0.470	1.674	1.471	0.800
		Tmr	114	30	30	27	27	0.470	1.674	1.471	0.800
		TOR	138	36	36	33	33	0.554	1.811	1.646	0.802
		TRR	144	36	36	36	36	0.405	1.041	1.095	0.806
	0.23	Tss	102	27	27	24	24	0.586	1.822	1.596	0.800
		Tinv	90	24	24	21	21	0.481	1.615	1.392	0.815
		Tmr	90	24	24	21	21	0.479	1.626	1.404	0.811
		TOR	102	27	27	24	24	0.573	1.727	1.578	0.801
		TRR	108	27	27	27	27	0.401	0.964	1.026	0.807
	0.25	Tss	90	24	24	21	21	0.621	1.786	1.562	0.811
		Tinv	72	18	18	18	18	0.337	1.537	1.325	0.801
		Tmr	72	18	18	18	18	0.335	1.532	1.324	0.800
		TOR	90	24	24	21	21	0.608	1.700	1.530	0.813
		TRR	90	24	24	21	21	0.405	0.923	0.983	0.804
(0.6,0.3,0.1)	0.2	Tss	138	36	36	33	33	0.568	1.920	1.692	0.801
		Tinv	120	30	30	30	30	0.496	1.746	1.530	0.802
		Tmr	120	30	30	30	30	0.496	1.746	1.530	0.802
		TOR	144	36	36	36	36	0.574	1.846	1.675	0.807
		TRR	150	39	39	36	36	0.455	1.058	1.040	0.805
	0.23	Tss	108	27	27	27	27	0.603	1.840	1.665	0.807
		Tinv	90	24	24	21	21	0.514	1.688	1.461	0.802
		Tmr	90	24	24	21	21	0.519	1.696	1.451	0.804
		TOR	108	27	27	27	27	0.590	1.770	1.640	0.806
		TRR	114	30	30	27	27	0.429	1.006	0.991	0.804
	0.25	Tss	90	24	24	21	21	0.636	1.841	1.622	0.804
		Tinv	78	21	21	18	18	0.555	1.669	1.432	0.805
		Tmr	78	21	21	18	18	0.532	1.671	1.436	0.804
		TOR	90	24	24	21	21	0.625	1.760	1.589	0.804
		TRR	96	24	24	24	24	0.376	0.941	0.949	0.801

Table 2. Sample sizes under $\delta_1 = \delta_2 = \delta_3 = \delta$ under k=1, (τ_1, τ_2)=(0.5, 0.5) and (r_1, r_2)=(1, 1).

$(\pi_1^0,\pi_2^0,\pi_3^0)$	δ	$(\Delta_1,\Delta_2,\Delta_3)$	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.4,0.2,0.1)	1.1	(0.27,0.23,0.15)	Tss	114	30	30	27	27	0.587	1.861	1.630	0.803
			Tinv	108	27	27	27	27	0.471	1.616	1.444	0.812
			Tmr	108	27	27	27	27	0.487	1.672	1.503	0.810
			TOR	114	30	30	27	27	0.579	1.786	1.606	0.803
			TRR	120	30	30	30	30	0.405	0.995	1.056	0.805
	1.25	(0.3,0.27,0.18)	Tss	84	21	21	21	21	0.524	1.703	1.543	0.802
			Tinv	78	21	21	18	18	0.482	1.587	1.354	0.813
			Tmr	78	21	21	18	18	0.521	1.595	1.401	0.809
			TOR	84	21	21	21	21	0.510	1.634	1.507	0.802
			TRR	90	24	24	21	21	0.422	0.925	0.983	0.805
	1.5	(0.35,0.33,0.23)	Tss	54	15	15	12	12	0.560	1.561	1.325	0.807
			Tinv	54	15	15	12	12	0.484	1.463	1.203	0.831
			Tmr	54	15	15	12	12	0.476	1.480	1.262	0.823
			TOR	54	15	15	12	12	0.520	1.476	1.308	0.804
			TRR	60	15	15	15	15	0.327	0.775	0.857	0.808
(0.6,0.3,0.1)	1.1	(0.22,0.26,0.15)	Tss	120	30	30	30	30	0.580	1.877	1.656	0.800
			Tinv	114	30	30	27	27	0.532	1.726	1.514	0.806
			Tmr	114	30	30	27	27	0.551	1.783	1.580	0.801
			TOR	126	33	33	30	30	0.605	1.835	1.634	0.812
			TRR	138	36	36	33	33	0.465	1.042	1.018	0.812
	1.25	(0.24,0.3,0.18)	Tss	96	24	24	24	24	0.610	1.835	1.640	0.806
			Tinv	90	24	24	21	21	0.547	1.696	1.466	0.814
			Tmr	90	24	24	21	21	0.566	1.726	1.516	0.809
			TOR	96	24	24	24	24	0.597	1.762	1.612	0.807
			TRR	108	27	27	27	27	0.389	0.969	0.989	0.808
	1.5	(0.27,0.36,0.23)	Tss	66	18	18	15	15	0.618	1.746	1.534	0.805
			Tinv	66	18	18	15	15	0.623	1.639	1.389	0.828
			Tmr	66	18	18	15	15	0.632	1.646	1.407	0.828
			TOR	66	18	18	15	15	0.596	1.645	1.471	0.807
			TRR	78	21	21	18	18	0.441	0.913	0.915	0.819

Table 3. Sample sizes under unequal differences and unequal odds ratios under k=1, (τ_1, τ_2)=(0.5, 0.5) and (r_1, r_2)=(1, 1).

(π_1^0,π_2^0,π_3^0)	(δ_1,δ_2,δ_3)	(Δ_1,Δ_2,Δ_3)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.4,0.2,0.1)	(1.25,1,0.81)	(0.3,0.2,0.1)	Tss	126	33	33	30	30	0.580	1.860	1.665	0.800
			Tinv	126	33	33	30	30	0.499	1.682	1.501	0.801
			Tmr	126	33	33	30	30	0.567	1.767	1.619	0.803
			TOR	126	33	33	30	30	0.573	1.797	1.636	0.803
			TRR	132	33	33	33	33	0.405	1.011	1.072	0.801
	(1.25,1.2,1.1)	(0.3,0.25,0.15)	Tss	96	24	24	24	24	0.527	1.782	1.591	0.806
			Tinv	90	24	24	21	21	0.510	1.611	1.399	0.813
			Tmr	90	24	24	21	21	0.544	1.645	1.490	0.808
			TOR	96	24	24	24	24	0.519	1.697	1.555	0.807
			TRR	102	27	27	24	24	0.422	0.969	1.024	0.802
	(1.5,1.15,0.93)	(0.35,0.24,0.12)	Tss	90	24	24	21	21	0.621	1.775	1.574	0.806
			Tinv	84	21	21	21	21	0.435	1.569	1.384	0.801
			Tmr	84	21	21	21	21	0.416	1.626	1.478	0.800
			TOR	90	24	24	21	21	0.605	1.696	1.540	0.807
			TRR	96	24	24	24	24	0.405	0.923	0.992	0.803
	(1.5,1.2,1.1)	(0.35,0.25,0.15)	Tss	84	21	21	21	21	0.551	1.767	1.574	0.807
			Tinv	78	21	21	18	18	0.524	1.591	1.338	0.818
			Tmr	78	21	21	18	18	0.574	1.653	1.439	0.810
			TOR	84	21	21	21	21	0.544	1.631	1.520	0.812
			TRR	84	21	21	21	21	0.365	0.878	0.964	0.800
(0.6,0.3,0.1)	(1.5,0.85,0.81)	(0.27,0.2,0.1)	Tss	144	36	36	36	36	0.579	1.932	1.692	0.806
			Tinv	138	36	36	33	33	0.559	1.758	1.532	0.809
			Tmr	138	36	36	33	33	0.617	1.875	1.654	0.805
			TOR	144	36	36	36	36	0.573	1.841	1.662	0.804
			TRR	162	42	42	39	39	0.455	1.064	1.048	0.803
	(1.5,0.9,1.1)	(0.27,0.21,0.15)	Tss	120	30	30	30	30	0.598	1.881	1.686	0.804
			Tinv	108	27	27	27	27	0.491	1.724	1.507	0.805
			Tmr	108	27	27	27	27	0.511	1.761	1.549	0.800
			TOR	120	30	30	30	30	0.583	1.831	1.650	0.803
			TRR	132	33	33	33	33	0.406	1.026	1.009	0.802
	(2.53,1.1,0.81)	(0.35,0.26,0.1)	Tss	96	24	24	24	24	0.641	1.837	1.640	0.809
			Tinv	90	24	24	21	21	0.587	1.713	1.463	0.807
			Tmr	90	24	24	21	21	0.674	1.821	1.585	0.803
			TOR	96	24	24	24	24	0.617	1.761	1.596	0.809
			TRR	108	27	27	27	27	0.429	0.983	0.983	0.802
	(2.53,1,1.1)	(0.35,0.24,0.15)	Tss	90	24	24	21	21	0.681	1.841	1.622	0.805
			Tinv	84	21	21	21	21	0.555	1.675	1.461	0.810
			Tmr	84	21	21	21	21	0.567	1.724	1.539	0.801
			TOR	90	24	24	21	21	0.650	1.759	1.590	0.801
			TRR	102	27	27	24	24	0.442	0.969	0.963	0.804

Table 4. Sample sizes under $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.4, 0.2, 0.1) , $(\Delta_1, \Delta_2, \Delta_3)$=(0.35, 0.24, 0.12), (τ_1, τ_2)=(0.5, 0.5) and (r_1, r_2)=(1, 1) for various k.

k	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
0.5	Tss	102	33	33	18	18	0.825	1.918	1.612	0.804
	Tinv	96	33	33	15	15	0.735	1.758	1.454	0.802
	Tmr	96	33	33	15	15	0.808	1.863	1.548	0.800
	TOR	102	33	33	18	18	0.810	1.851	1.583	0.804
	TRR	114	39	39	18	18	0.584	1.057	0.969	0.807
1	Tss	108	27	27	27	27	0.622	1.877	1.681	0.802
	Tinv	102	27	27	24	24	0.540	1.705	1.481	0.805
	Tmr	102	27	27	24	24	0.578	1.776	1.590	0.801
	TOR	108	27	27	27	27	0.605	1.794	1.639	0.803
	TRR	120	30	30	30	30	0.413	0.993	0.999	0.804
2	Tss	120	21	21	39	39	0.370	1.817	1.717	0.812
	Tinv	114	18	18	39	39	0.276	1.652	1.549	0.805
	Tmr	114	18	18	39	39	0.281	1.688	1.659	0.803
	TOR	120	21	21	39	39	0.351	1.730	1.675	0.812
	TRR	132	21	21	45	45	0.218	0.913	1.032	0.804

Table 5. Sample sizes under $\Delta_1 = \Delta_2 = \Delta_3 = \Delta$ under k=1, (τ_1, τ_2)=(0.25, 0.25) and (r_1, r_2)=(2, 1).

$(\pi_1^0,\pi_2^0,\pi_3^0)$	Δ	(r_1,r_2)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.4,0.2,0.1)	0.2	(1,1)	Tss	126	33	33	30	30	0.157	2.115	1.592	0.800
			Tinv	108	27	27	27	27	0.000	1.888	1.427	0.803
			Tmr	108	27	27	27	27	0.000	1.888	1.427	0.803
			TOR	126	33	33	30	30	0.154	2.006	1.569	0.801
			TRR	126	33	33	30	30	0.179	1.090	1.026	0.806
		(2,1)	Tss	120	24	48	24	24	0.174	1.853	1.441	0.805
			Tinv	96	21	39	18	18	0.052	1.673	1.218	0.801
			Tmr	96	21	39	18	18	0.052	1.673	1.218	0.801
			TOR	120	24	48	24	24	0.173	1.768	1.416	0.803
			TRR	120	24	48	24	24	0.124	0.949	0.935	0.803
	0.23	(1,1)	Tss	96	24	24	24	24	0.000	2.038	1.532	0.804
			Tinv	78	21	21	18	18	0.000	1.811	1.300	0.801
			Tmr	78	21	21	18	18	0.000	1.811	1.300	0.801
			TOR	96	24	24	24	24	0.150	1.894	1.498	0.807
			TRR	96	24	24	24	24	0.110	0.971	0.965	0.806
		(2,1)	Tss	87	18	33	18	18	0.038	1.741	1.334	0.800
			Tinv	72	15	27	15	15	-0.052	1.514	1.115	0.802
			Tmr	72	15	27	15	15	-0.052	1.514	1.115	0.802
			TOR	87	18	33	18	18	0.038	1.628	1.302	0.805
			TRR	90	18	36	18	18	0.000	0.851	0.849	0.802
	0.25	(1,1)	Tss	78	21	21	18	18	0.166	1.945	1.452	0.803
			Tinv	66	18	18	15	15	0.000	1.772	1.250	0.800
			Tmr	66	18	18	15	15	0.000	1.747	1.263	0.800
			TOR	78	21	21	18	18	0.157	1.824	1.420	0.804
			TRR	78	21	21	18	18	0.116	0.926	0.905	0.800
		(2,1)	Tss	72	15	27	15	15	0.025	1.630	1.234	0.805
			Tinv	60	12	24	12	12	-0.134	1.320	0.969	0.808
			Tmr	60	12	24	12	12	-0.134	1.320	0.969	0.808
			TOR	72	15	27	15	15	0.025	1.519	1.218	0.804
			TRR	75	15	30	15	15	0.000	0.778	0.778	0.801

Table 5. Continue.

$(\pi_1^0,\pi_2^0,\pi_3^0)$	Δ	(r_1,r_2)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.6,0.3,0.1)	0.2	(1,1)	Tss	126	33	33	30	30	0.163	2.158	1.596	0.804
			Tinv	108	27	27	27	27	0.010	2.028	1.435	0.803
			Tmr	108	27	27	27	27	0.000	2.016	1.448	0.800
			TOR	126	33	33	30	30	0.254	2.048	1.575	0.805
			TRR	132	33	33	33	33	0.169	1.104	0.985	0.800
		(2,1)	Tss	126	27	51	24	24	0.240	2.063	1.531	0.804
			Tinv	105	21	42	21	21	0.066	1.795	1.316	0.807
			Tmr	105	21	42	21	21	0.047	1.774	1.323	0.806
			TOR	126	27	51	24	24	0.237	1.961	1.502	0.806
			TRR	138	27	57	27	27	0.178	1.050	0.949	0.801
	0.23	(1,1)	Tss	96	24	24	24	24	0.166	2.077	1.586	0.801
			Tinv	84	21	21	21	21	0.000	1.932	1.398	0.802
			Tmr	84	21	21	21	21	0.000	1.911	1.403	0.803
			TOR	96	24	24	24	24	0.156	1.989	1.547	0.801
			TRR	108	27	27	27	27	0.185	1.052	0.963	0.807
		(2,1)	Tss	93	18	39	18	18	0.137	1.927	1.413	0.806
			Tinv	78	15	33	15	15	0.045	1.673	1.195	0.807
			Tmr	78	15	33	15	15	0.037	1.664	1.188	0.808
			TOR	93	18	39	18	18	0.135	1.782	1.381	0.809
			TRR	105	21	42	21	21	0.137	0.953	0.893	0.810
	0.25	(1,1)	Tss	84	21	21	21	21	0.175	2.020	1.549	0.809
			Tinv	72	18	18	18	18	0.000	1.870	1.360	0.804
			Tmr	72	18	18	18	18	0.000	1.908	1.373	0.801
			TOR	84	21	21	21	21	0.173	1.905	1.518	0.810
			TRR	90	24	24	21	21	0.192	1.014	0.920	0.805
		(2,1)	Tss	75	15	30	15	15	0.128	1.835	1.354	0.800
			Tinv	63	12	27	12	12	-0.018	1.595	1.089	0.802
			Tmr	63	12	27	12	12	-0.018	1.595	1.089	0.802
			TOR	75	15	30	15	15	0.125	1.650	1.316	0.802
			TRR	87	18	33	18	18	0.186	0.908	0.863	0.803

Table 6. Sample sizes under $\delta_1 = \delta_2 = \delta_3 = \delta$, k=1, (τ_1, τ_2)=(0.25, 0.25) and (r_1, r_2)=(2, 1)

$(\pi_1^0,\pi_2^0,\pi_3^0)$	δ	$(\Delta_1,\Delta_2,\Delta_3)$	(r_1,r_2)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.4,0.2,0.1)	1.1	(0.27,0.23,0.15)	(1,1)	Tss	102	27	27	24	24	0.158	2.058	1.551	0.802
				Tinv	96	24	24	24	24	0.028	1.852	1.371	0.808
				Tmr	96	24	24	24	24	0.014	1.903	1.437	0.802
				TOR	102	27	27	24	24	0.156	1.946	1.520	0.804
				TRR	108	27	27	27	27	0.110	1.014	0.993	0.804
			(2,1)	Tss	102	21	39	21	21	0.172	1.847	1.422	0.807
				Tinv	87	18	33	18	18	0.014	1.576	1.178	0.805
				Tmr	87	18	33	18	18	0.004	1.632	1.234	0.801
				TOR	96	21	39	18	18	0.178	1.713	1.349	0.802
				TRR	102	21	39	21	21	0.137	0.904	0.902	0.804
	1.25	(0.3,0.27,0.18)	(1,1)	Tss	78	21	21	18	18	0.175	1.955	1.466	0.807
				Tinv	72	18	18	18	18	0.000	1.760	1.289	0.808
				Tmr	72	18	18	18	18	0.000	1.806	1.331	0.809
				TOR	78	21	21	18	18	0.173	1.842	1.438	0.807
				TRR	84	21	21	21	21	0.123	0.949	0.948	0.810
			(2,1)	Tss	72	15	27	15	15	0.028	1.647	1.258	0.807
				Tinv	60	12	24	12	12	-0.125	1.322	0.971	0.804
				Tmr	60	12	24	12	12	-0.157	1.326	0.984	0.803
				TOR	72	15	27	15	15	0.027	1.543	1.227	0.809
				TRR	75	15	30	15	15	0.000	0.778	0.784	0.807
	1.5	(0.35,0.33,0.23)	(1,1)	Tss	54	15	15	12	12	0.203	1.780	1.304	0.825
				Tinv	48	12	12	12	12	0.000	1.577	1.160	0.805
				Tmr	48	12	12	12	12	0.000	1.582	1.176	0.807
				TOR	54	15	15	12	12	0.196	1.650	1.272	0.827
				TRR	54	15	15	12	12	0.141	0.828	0.806	0.805
			(2,1)	Tss	45	9	18	9	9	0.000	1.197	0.885	0.823
				Tinv	33	6	15	6	6	-0.314	0.803	0.427	0.809
				Tmr	33	6	15	6	6	-0.309	0.800	0.437	0.803
				TOR	45	9	18	9	9	0.000	1.140	0.870	0.820
				TRR	45	9	18	9	9	0.000	0.632	0.564	0.805

Table 6. Continue.

$(\pi_1^0,\pi_2^0,\pi_3^0)$	δ	$(\Delta_1,\Delta_2,\Delta_3)$	(r_1,r_2)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.6,0.3,0.1)	1.1	(0.22,0.26,0.15)	(1,1)	Tss	114	30	30	27	27	0.28	2.155	1.612	0.810
				Tinv	102	27	27	24	24	0.095	1.989	1.436	0.801
				Tmr	108	27	27	27	27	0.107	2.004	1.517	0.810
				TOR	114	30	30	27	27	0.269	2.042	1.575	0.811
				TRR	120	30	30	30	30	0.174	1.086	0.970	0.800
			(2,1)	Tss	111	24	45	21	21	0.256	2.048	1.500	0.811
				Tinv	96	21	39	18	18	0.084	1.785	1.279	0.803
				Tmr	102	21	39	21	21	0.089	1.827	1.374	0.804
				TOR	108	21	45	21	21	0.125	1.839	1.449	0.800
				TRR	123	24	51	24	24	0.18	1.005	0.920	0.805
	1.25	(0.24,0.3,0.18)	(1,1)	Tss	84	21	21	21	21	0.17	2.020	1.522	0.801
				Tinv	78	21	21	18	18	0.113	1.935	1.376	0.801
				Tmr	84	21	21	21	21	0.111	1.952	1.450	0.812
				TOR	84	21	21	21	21	0.165	1.905	1.492	0.801
				TRR	96	24	24	24	24	0.185	1.014	0.942	0.802
			(2,1)	Tss	81	18	33	15	15	0.167	1.930	1.390	0.802
				Tinv	75	15	30	15	15	0.059	1.678	1.210	0.810
				Tmr	75	15	30	15	15	0.059	1.699	1.234	0.807
				TOR	81	18	33	15	15	0.202	1.790	1.372	0.805
				TRR	93	18	39	18	18	0.153	0.929	0.849	0.802
	1.5	(0.27,0.36,0.23)	(1,1)	Tss	60	15	15	15	15	0.192	1.857	1.426	0.802
				Tinv	60	15	15	15	15	0.075	1.823	1.330	0.821
				Tmr	60	15	15	15	15	0.07	1.828	1.327	0.822
				TOR	60	15	15	15	15	0.185	1.749	1.392	0.803
				TRR	66	18	18	15	15	0.208	0.925	0.848	0.801
			(2,1)	Tss	57	12	21	12	12	0.142	1.684	1.277	0.813
				Tinv	48	9	21	9	9	0.062	1.400	0.950	0.819
				Tmr	48	9	21	9	9	0.049	1.374	0.948	0.816
				TOR	57	12	21	12	12	0.138	1.544	1.229	0.819
				TRR	63	12	27	12	12	0.117	0.783	0.731	0.814

Table 7. Sample sizes under unequal differences and unequal odds ratios under k=1, (τ_1,τ_2)=(0.25, 0.25) and (r_1,r_2)=(2, 1).

$(\pi_1^0,\pi_2^0,\pi_3^0)$	$(\delta_1,\delta_2,\delta_3)$	$(\Delta_1,\Delta_2,\Delta_3)$	(r_1,r_2)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.4,0.2,0.1)	(1.25,1,0.81)	(0.3,0.2,0.1)	(1,1)	Tss	114	30	30	27	27	0.157	2.098	1.567	0.806
				Tinv	114	30	30	27	27	0.100	1.926	1.414	0.802
				Tmr	114	30	30	27	27	0.130	1.987	1.514	0.807
				TOR	114	30	30	27	27	0.161	1.979	1.541	0.804
				TRR	120	30	30	30	30	0.113	1.052	1.016	0.805
			(2,1)	Tss	108	21	45	21	21	0.027	1.773	1.387	0.801
				Tinv	105	21	42	21	21	0.058	1.639	1.252	0.801
				Tmr	105	21	42	21	21	0.060	1.715	1.347	0.802
				TOR	108	21	45	21	21	0.113	1.681	1.359	0.802
				TRR	120	24	48	24	24	0.130	0.953	0.940	0.807
	(1.25,1.2,1.1)	(0.3,0.25,0.15)	(1,1)	Tss	84	21	21	21	21	0.000	1.955	1.478	0.802
				Tinv	78	21	21	18	18	0.031	1.812	1.309	0.803
				Tmr	78	21	21	18	18	0.035	1.847	1.361	0.800
				TOR	84	21	21	18	18	0.030	1.860	1.375	0.800
				TRR	90	24	24	21	21	0.131	0.970	0.948	0.808
			(2,1)	Tss	78	15	33	15	15	-0.023	1.564	1.213	0.804
				Tinv	72	15	27	15	15	-0.042	1.471	1.109	0.806
				Tmr	72	15	27	15	15	-0.034	1.543	1.139	0.806
				TOR	78	15	33	15	15	-0.021	1.455	1.180	0.807
				TRR	87	18	33	18	18	0.061	0.867	0.863	0.805
	(1.5,1.15,0.93)	(0.35,0.24,0.12)	(1,1)	Tss	84	21	21	21	21	0.173	1.945	1.497	0.808
				Tinv	78	21	21	18	18	0.068	1.812	1.306	0.807
				Tmr	78	21	21	18	18	0.088	1.894	1.391	0.806
				TOR	84	21	21	21	21	0.169	1.842	1.464	0.813
				TRR	90	24	24	21	21	0.200	0.971	0.957	0.812
			(2,1)	Tss	75	15	30	15	15	0.000	1.585	1.235	0.806
				Tinv	72	15	27	15	15	-0.025	1.471	1.109	0.809
				Tmr	72	15	27	15	15	-0.001	1.563	1.167	0.812
				TOR	75	15	30	15	15	0.000	1.454	1.199	0.804
				TRR	78	15	33	15	15	0.014	0.781	0.778	0.801
	(1.5,1.2,1.1)	(0.35,0.25,0.15)	(1,1)	Tss	78	21	21	18	18	0.181	1.839	1.440	0.822
				Tinv	72	18	18	18	18	0.026	1.754	1.300	0.819
				Tmr	72	18	18	18	18	0.031	1.818	1.375	0.812
				TOR	78	21	21	18	18	0.179	1.841	1.439	0.822
				TRR	78	21	21	18	18	0.131	0.949	0.913	0.806
			(2,1)	Tss	63	12	27	12	12	-0.063	1.396	1.064	0.801
				Tinv	60	12	24	12	12	-0.100	1.328	0.979	0.803
				Tmr	60	12	24	12	12	-0.108	1.331	1.009	0.805
				TOR	63	12	27	12	12	-0.085	1.286	1.035	0.803
				TRR	72	15	27	15	15	0.046	0.798	0.789	0.807

Table 7. Continue.

$(\pi_1^0,\pi_2^0,\pi_3^0)$	$(\delta_1,\delta_2,\delta_3)$	$(\Delta_1,\Delta_2,\Delta_3)$	(r_1,r_2)	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
(0.6,0.3,0.1)	(1.5,0.85,0.81)	(0.27,0.2,0.1)	(1,1)	Tss	132	33	33	33	33	0.255	2.195	1.627	0.811
				Tinv	126	33	33	30	30	0.136	1.995	1.477	0.802
				Tmr	126	33	33	30	30	0.159	2.112	1.586	0.800
				TOR	132	33	33	33	33	0.157	2.079	1.617	0.806
				TRR	144	36	36	36	36	0.169	1.131	0.996	0.803
			(2,1)	Tss	126	27	51	24	24	0.246	2.063	1.542	0.802
				Tinv	123	24	51	24	24	0.098	1.810	1.354	0.804
				Tmr	123	24	51	24	24	0.120	1.904	1.463	0.802
				TOR	126	27	51	24	24	0.246	2.063	1.542	0.802
				TRR	147	30	57	30	30	0.158	1.090	0.962	0.803
	(1.5,0.9,1.1)	(0.27,0.21,0.15)	(1,1)	Tss	96	24	24	24	24	0.166	2.097	1.574	0.804
				Tinv	90	24	24	21	21	0.134	1.927	1.424	0.803
				Tmr	90	24	24	21	21	0.146	1.978	1.478	0.801
				TOR	96	24	24	24	24	0.164	1.989	1.536	0.803
				TRR	108	27	27	27	27	0.185	1.052	0.963	0.806
			(2,1)	Tss	93	18	39	18	18	0.127	1.900	1.421	0.804
				Tinv	87	18	33	18	18	0.083	1.756	1.273	0.810
				Tmr	87	18	33	18	18	0.084	1.800	1.337	0.807
				TOR	93	18	39	18	18	0.130	1.756	1.389	0.809
				TRR	105	21	42	21	21	0.133	0.972	0.888	0.805
	(2.53,1.1,0.81)	(0.35,0.26,0.1)	(1,1)	Tss	96	24	24	24	24	0.169	2.083	1.565	0.800
				Tinv	96	24	24	24	24	0.138	1.933	1.440	0.808
				Tmr	96	24	24	24	24	0.142	2.037	1.524	0.808
				TOR	96	24	24	24	24	0.163	1.983	1.525	0.801
				TRR	108	27	26	27	26	0.185	1.052	0.949	0.802
			(2,1)	Tss	93	18	39	18	18	0.125	1.900	1.421	0.801
				Tinv	90	18	36	18	18	0.085	1.755	1.263	0.813
				Tmr	90	18	36	18	18	0.094	1.800	1.360	0.804
				TOR	93	18	39	18	18	0.126	1.756	1.389	0.803
				TRR	111	24	45	21	21	0.224	1.008	0.910	0.807
	(2.53,1,1.1)	(0.35,0.24,0.15)	(1,1)	Tss	90	24	24	21	21	0.306	2.104	1.569	0.810
				Tinv	84	21	21	21	21	0.112	1.935	1.411	0.808
				Tmr	84	21	21	21	21	0.112	1.944	1.476	0.805
				TOR	90	24	24	21	21	0.289	2.005	1.540	0.809
				TRR	96	24	24	24	24	0.179	1.014	0.930	0.801
			(2,1)	Tss	87	18	33	18	18	0.165	1.943	1.447	0.811
				Tinv	75	15	30	15	15	0.043	1.680	1.197	0.805
				Tmr	78	15	33	15	15	0.070	1.716	1.257	0.805
				TOR	81	18	33	15	15	0.154	1.789	1.355	0.802
				TRR	96	21	39	18	18	0.184	0.958	0.865	0.806

Table 8. Sample sizes under $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.4, 0.2, 0.1), $(\Delta_1, \Delta_2, \Delta_3)$=(0.35, 0.24, 0.12), (τ_1,τ_2)=(0.25, 0.25) and (r_1, r_2)=(2, 1) for various k.

k	Method	N	M_1^c	M_1^e	M_2^c	M_2^e	a_1	b_1	b_2	Power
0.5	Tss	72	18	36	9	9	0.212	1.691	1.118	0.802
	Tinv	69	18	33	9	9	0.155	1.577	1.021	0.806
	Tmr	69	18	33	9	9	0.197	1.686	1.088	0.800
	TOR	72	18	36	9	9	0.209	1.571	1.099	0.800
	TRR	81	21	42	9	9	0.255	0.898	0.759	0.812
1	Tss	75	15	30	15	15	0.000	1.599	1.230	0.805
	Tinv	72	15	27	15	15	-0.019	1.471	1.094	0.815
	Tmr	72	15	27	15	15	0.003	1.560	1.190	0.804
	TOR	75	15	30	15	15	0.000	1.480	1.190	0.804
	TRR	81	18	33	15	15	0.106	0.862	0.820	0.801
2	Tss	78	12	24	21	21	-0.252	1.412	1.301	0.811
	Tinv	72	9	21	21	21	-0.379	1.117	1.090	0.808
	Tmr	72	9	21	21	21	-0.358	1.139	1.152	0.807
	TOR	78	12	24	21	21	-0.236	1.301	1.278	0.810
	TRR	84	12	24	24	24	-0.179	0.711	0.852	0.809

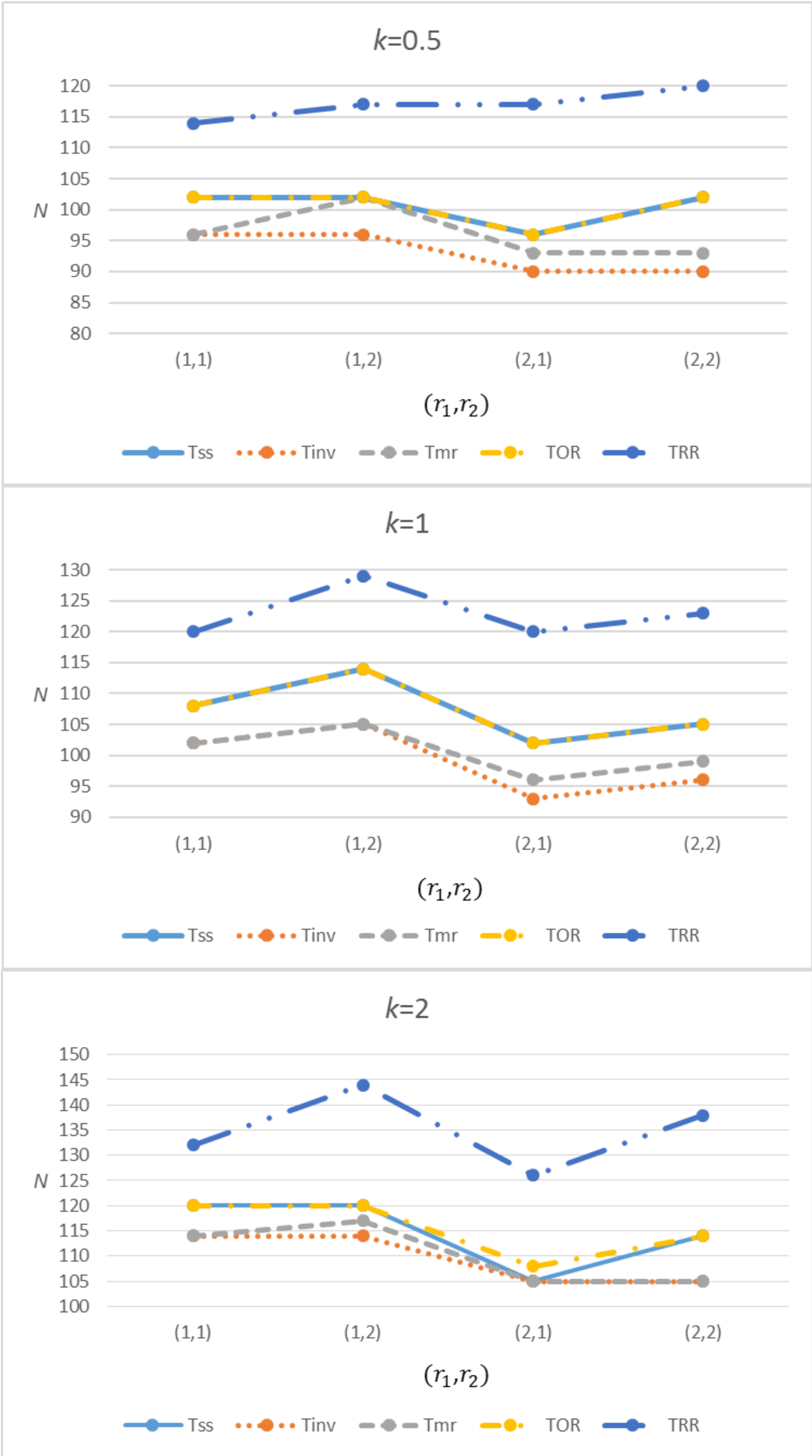

Figure 1. Total sample sizes under $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.4, 0.2, 0.1), $(\Delta_1, \Delta_2, \Delta_3)$=(0.35, 0.24, 0.12) and $\tau_1 = \tau_2$ =0.5 for *k*= (0.5, 1, 2) and various (r_1, r_2).

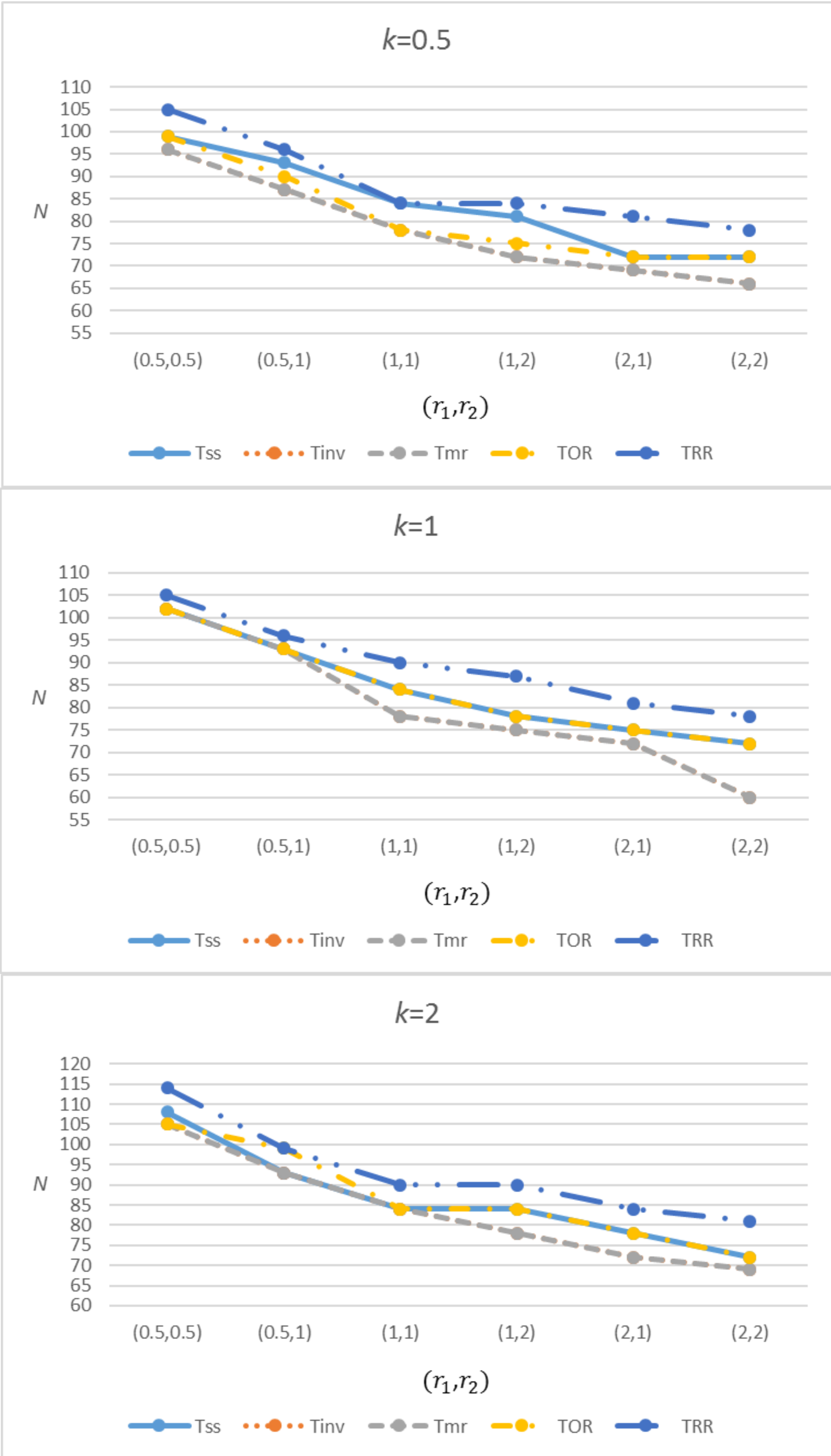

Figure 2. Total sample sizes under $(\pi_1^0, \pi_2^0, \pi_3^0)$=(0.4, 0.2, 0.1), $(\Delta_1, \Delta_2, \Delta_3)$=(0.35, 0.24, 0.12) and $\tau_1 = \tau_2$ =0.25 for *k*= (0.5, 1, 2) and various (r_1, r_2).