Published in final edited form as:

Stat Med. 2017 February 20; 36(4): 581–591. doi:10.1002/sim.7168.

Inference and Sample Size Calculation for Clinical Trials with Incomplete Observations of Paired Binary Outcomes[†]

Song Zhang^a, Jing Cao^b, and Chul Ahn^a

^aDepartment of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX

^bDepartment of Statistical Science, Southern Methodist University, Dallas, TX

Abstract

We investigate the estimation of intervention effect and sample size determination for experiments where subjects are supposed to contribute paired binary outcomes with some incomplete observations. We propose a hybrid estimator to appropriately account for the mixed nature of observed data: paired outcomes from those who contribute complete pairs of observations and unpaired outcomes from those who contribute either pre- or post-intervention outcomes. We theoretically prove that if incomplete data are evenly distributed between the pre- and post-intervention periods, the proposed estimator will always be more efficient than the traditional estimator. A numerical research shows that when the distribution of incomplete data is unbalanced, the proposed estimator will be superior when there is moderate-to-strong positive within-subject correlation. We further derive a closed-form sample size formula to help researchers determine how many subjects need to be enrolled in such studies. Simulation results suggest that the calculated sample size maintain the empirical power and type I error under various design configurations. We demonstrate the proposed method using a real application example.

Keywords

Sample Size; Paired; Binary Outcomes; incomplete

1. Introduction

Paired experimental design is frequently adopted in scientific studies. In many cases, the outcome of primary interest is binary. For example, in radiology, researchers perform two imaging procedures on the same group of patients to compare the rates of correct diagnosis between procedures. In social and behavioral sciences, pre-post studies are conducted such that responses from the same group of subjects before and after an intervention are compared to evaluate the effectiveness of an intervention. In the above two examples, each experimental unit contributes a pair of outcome measurements. Paired outcome measurements can also arise from matched case-control studies or matched randomized

[†]Please ensure that you use the most up to date class file, available from the SIM Home Page at www.interscience.wiley.com/jpages/0277-6715

*Correspondence to: Chul Ahn, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX. chul.ahn@utsouthwestern.edu.

trials, where subjects in two comparison groups are matched by a set of characteristics potentially related to the outcome.

In practice, researchers conducting paired experiments often encounter the missing data problem due to patient refusal, lost to follow-up, equipment failure or unavailability, etc. In such scenarios researchers are presented with a “mixture” type of data: paired (dependent) outcomes from subjects who contribute complete data and unpaired (independent) outcomes from those who contribute incomplete data. For example, non-invasive diagnostic tests of ultrasonography (US), magnetic resonance (MR) and echocolonoscopy (EC) were evaluated for their ability to provide correct diagnosis of endometriosis, but a significant portion of patients have missing values for MR or EC due to unavailability of the corresponding equipments at the occasion of evaluation [1]. Missing data can also be caused by protocol changes during the trial. Campbell *et al.* [2] described a controlled clinical trial on the safety and effectiveness of a cardiovascular device against the standard medical therapy, where in the middle of the trial, a protocol change in the assessment of primary outcome leads to missing baseline measurements in 60% of patients. Sometimes the “mixture” type of data arises from the experimental design. In a pre-post survey on 1000 subjects to evaluate the impact of a 3-month radio campaign on people's attitude toward colorectal cancer screening. Expecting that 20% of the subjects will be lost to follow-up at the end of the campaign, researchers plan to include 200 new subjects in the post-campaign survey. Thus a total of 1200 unique subjects will be involved, 200 pre-campaign only, 200 post-campaign only, and 800 both.

If every subject contributes a complete pair of observations, then the McNemar's test is a straightforward choice to detect difference. On the other hand, if the outcomes are obtained from two independent groups of subjects, then the Chi-square test will be appropriate. The sample size calculation for the McNemar's test or the Chi-square test has been well established [3, 4, 5]. Recently there has been some development in statistical inference based on paired binary outcomes with missing data. Based on the large sample theory, various estimators were proposed along with the confidence intervals for the intervention effect [6, 7, 8]. Shih [9] proposed maximum likelihood estimation and likelihood ratio test with incomplete pairs. They employed an EM algorithm in computation. Choi and Stablein [10] investigated the effect of non-random missing data mechanisms. Tang and Tang [11] proposed an exact-test approach that compares two paired proportions with incomplete data. Tang *et al.* [12] investigated exact and approximate unconditional confidence intervals for proportional difference in the presence of incomplete data. Li et al. [13] developed a hypothesis testing method for a simple ordering in incomplete contingency tables and Li *et al.* [14] constructed confidence intervals for rate ratio in matched-pair studies with incomplete data. Despite the extensive application of paired experimental design, there has been limited research on sample size determination for such studies accounting for the impact of missing data.

The rest of the paper is arranged as follows. In Section 2 we review the existing methods and present a new test statistic which combines information from paired and unpaired observations. Furthermore, closed-form sample size formulas are presented for paired experimental design with incomplete binary outcomes. In Section 3 we conduct simulation

studies to examine the performance of sample size formulas. In Section 4 we demonstrate the sample size method using a real application example. Finally, we discuss limitations and possible extensions in Section 5.

2. Statistical Inference and Sample Size Calculation

We introduce the following notations. Let Y_{it} be a binary outcome (1 for Yes and 0 for No) from the i th ($i = 1, \dots, n$) subject under treatment t ($t = 0, 1$). Here we generically use treatment to denote research categories: case/control, MR/EC, pre-/post-interventions, etc. For paired outcomes (Y_{i0}, Y_{i1}), we define joint probabilities

$$h_{uv} = P(Y_{i0} = u, Y_{i1} = v) \quad \text{for } u = 0, 1 \quad \text{and} \quad v = 0, 1.$$

Then we have marginal probabilities $p_0 = P(Y_{i0} = 1) = h_{10} + h_{11}$ and $p_1 = P(Y_{i1} = 1) = h_{01} + h_{11}$. The relationship between joint and marginal probabilities are illustrated in Table 1. We also define $\rho = \text{Corr}(Y_{i0}, Y_{i1})$ to be the within-subject correlation coefficient. It can be shown that

$$\rho = \frac{h_{11} - p_0 p_1}{\sqrt{p_0(1-p_0)p_1(1-p_1)}}. \quad (1)$$

For binary variables, there is a range for the valid values of within-subject correlation, $\rho \in [\rho_L, \rho_U]$, determined by the marginal probabilities:

$$\begin{aligned} \rho_L &= \max \left\{ -\sqrt{\frac{p_0 p_1}{(1-p_0)(1-p_1)}}, -\sqrt{\frac{(1-p_0)(1-p_1)}{p_0 p_1}} \right\}, \\ \rho_U &= \min \left\{ \sqrt{\frac{p_0(1-p_1)}{p_1(1-p_0)}}, \sqrt{\frac{p_1(1-p_0)}{p_0(1-p_1)}} \right\}. \end{aligned}$$

Details of the derivation can be found in Chaganty and Joe [15]. Equation (1) also suggests that there is a one-to-one correspondence between (p_0, p_1, h_{11}) and (p_0, p_1, ρ) , and they fully characterize the statistical properties of (Y_{i0}, Y_{i1}) . In the rest of this paper we will use (p_0, p_1, ρ) and (p_0, p_1, h_{11}) exchangeably in the description of paired outcomes. The treatment effect is denoted by $\tau = p_1 - p_0 = h_{01} - h_{10}$. We are interested in testing the null hypothesis $H_0: p_1 - p_0 = 0$.

Without loss of generality, we define $n = m_1 + m_2 + m_3$ and number the subjects (from 1 to n) in such a fashion that the first m_1 subjects only contribute outcomes under treatment 0, the next m_2 subjects contribute outcomes under both treatments, and the last m_3 subjects only contribute outcomes under treatment 1. We use $\mathbf{Y}_1 = \{Y_{i0}, i = 1, \dots, m_1\}$, $\mathbf{Y}_2 = \{(Y_{i0}, Y_{i1}), i = m_1 + 1, \dots, m_1 + m_2\}$, and $\mathbf{Y}_3 = \{Y_{i1}, i = m_1 + m_2 + 1, \dots, n\}$ to denote the collection of outcomes from the three groups. We also define $r = m_1/n$ and $s = m_3/n$ to be the proportions of subjects with partial outcomes. In the following discussion, we assume a missing completely at random mechanism. That is, the occurrence of missing values is independent of the observed or unobserved outcomes. Note that the setting described above

is equivalent to that assumed by Tang *et al.* [16], which stated that paired observations ($\mathbf{Y}_2$) follow a multinomial distribution with probabilities ($h_{00}, h_{01}, h_{10}, h_{11}$), and incomplete observations ($\mathbf{Y}_1$ and $\mathbf{Y}_3$) follow independent binomial distributions with probabilities p_0 and p_1 .

2.1. Existing Methods to Estimate Intervention Effect

The general idea of existing approaches to estimating the treatment effect proceeds as follows. 1) Using only complete pairs of observations from $\mathbf{Y}_2$ to obtain estimators of p_0 and p_1 , denoted by $\hat{p}_0^{(P)}$ and $\hat{p}_1^{(P)}$. Then use unpaired observations from $\mathbf{Y}_1$ and $\mathbf{Y}_3$ to obtain estimators $\hat{p}_0^{(U)}$ and $\hat{p}_1^{(U)}$, respectively. Here superscripts (P) and (U) indicate whether the estimators are based on paired or unpaired observations. 2) Because $\hat{p}_t^{(P)}$ and $\hat{p}_t^{(U)}$ are both unbiased estimators of the true response rates p_t ($t = 0, 1$), a class of hybrid estimators can be constructed as the weighted average of $\hat{p}_t^{(P)}$ and $\hat{p}_t^{(U)}$,

$$\hat{p}_t^{(H)} = \phi_t \hat{p}_t^{(P)} + (1 - \phi_t) \hat{p}_t^{(U)},$$

with weight $0 \leq \phi_t \leq 1$. Here the superscript (H) indicates the hybrid nature of the estimator.

3) A class of unbiased estimator for treatment effect is $\hat{\Delta}_1 = \hat{p}_1^{(H)} - \hat{p}_0^{(H)}$.

Specifically, the aforementioned estimators of p_0 and p_1 are

$$\begin{aligned}\hat{p}_t^{(P)} &= \sum_{i=m_1+1}^{m_1+m_2} Y_{it} / m_2 \quad \text{based on } \mathbf{Y}_2 \quad \text{for } t=0,1; \\ \hat{p}_0^{(U)} &= \sum_{i=1}^{m_1} Y_{i0} / m_1 \quad \text{based on } \mathbf{Y}_1; \\ \hat{p}_1^{(U)} &= \sum_{i=m_1+m_2+1}^n Y_{i1} / m_3 \quad \text{based on } \mathbf{Y}_3.\end{aligned}$$

A traditionally adopted weighting scheme [6, 7, 8] is to set $\phi_0 = m_2 / (m_1 + m_2)$ and $\phi_1 = m_2 / (m_2 + m_3)$. Under this weighting scheme, the hybrid estimator can be simplified as

$$\hat{p}_0^{(H^*)} = \frac{\sum_{i=1}^{m_1+m_2} Y_{i0}}{m_1+m_2}, \quad \hat{p}_1^{(H^*)} = \frac{\sum_{i=m_1+1}^n Y_{i1}}{m_2+m_3}, \quad (2)$$

and $\hat{\Delta}_1^* = \hat{p}_1^{(H^*)} - \hat{p}_0^{(H^*)}$.

Fact 1—Under weights $\phi_0 = m_2 / (m_1 + m_2)$ and $\phi_1 = m_2 / (m_2 + m_3)$, $\sqrt{n} (\hat{\Delta}_1^* - \Delta)$ approximately follows a normal distribution with mean 0 and variance

$$\sigma_1^2 = \frac{p_0(1-p_0)}{1-s} + \frac{p_1(1-p_1)}{1-r} - \frac{2(1-r-s)(h_{11} - p_0p_1)}{(1-s)(1-r)}. \quad (3)$$

The expression of σ_1^2 has been investigated by many researchers [6, 7, 8]. For readers' reference, we include the derivation details in Appendix A.

2.2. The Proposed Approach

The pre- and post-intervention outcomes (Y_{i0}, Y_{i1}) are observed as pairs in $\mathbf{Y}_2$. In constructing the hybrid estimator in (2), however, they are used as if they were independently observed. i.e., Y_{i0} and Y_{i1} from $\mathbf{Y}_2$ are used separately in $\hat{p}_0^{(H*)}$ and $\hat{p}_1^{(H*)}$. As a result, although $\hat{\Delta}_1^*$ remains an unbiased estimator, it might lose efficiency. In the following, we propose a different hybrid estimator which directly estimates the treatment effect (instead of estimating p_0 and p_1 first and then taking the difference) based on paired and unpaired observations, denoted by $\hat{\Delta}^{(P)}$ and $\hat{\Delta}^{(U)}$. The advantage of $\hat{\Delta}^{(P)}$ is that it accounts for the pairing of observations in $\mathbf{Y}_2$. Then a new class of hybrid estimator can be defined as the weighted sum of $\hat{\Delta}^{(P)}$ and $\hat{\Delta}^{(U)}$, $\hat{\Delta}_2 = w\hat{\Delta}^{(P)} + (1-w)\hat{\Delta}^{(U)}$, with weight $0 \leq w \leq 1$. We further investigate the optimal weighting scheme that minimizes the variance of $\hat{\Delta}_2$.

$$\tau_1^2 = p_1(1-p_1)/s + p_0(1-p_0)/r. \quad (4)$$

The estimator of the treatment effect based on independent observations ($\mathbf{Y}_1$ and $\mathbf{Y}_3$) is

$\hat{\Delta}^{(U)} = \hat{p}_1^{(U)} - \hat{p}_0^{(U)}$. The large sample theory suggests that $\sqrt{n}(\hat{\Delta}^{(U)} - \Delta)$ is approximately normally distributed with mean 0 and variance

The estimator of the treatment effect based on paired observations ($\mathbf{Y}_2$) is constructed using

within-subject difference $d_i = Y_{i1} - Y_{i0}$, $\hat{\Delta}^{(P)} = \sum_{i=m_1+1}^{m_1+m_2} d_i / m_2$, which takes into account potential within-subject correlation. [4] has shown that $\text{Var}(d_i) = h_{01} + h_{10} - (h_{01} - h_{10})^2$.

Thus the approximate distribution of $\sqrt{n}(\hat{\Delta}^{(P)} - \Delta)$ is normal with mean 0 and variance

$$\tau_2^2 = \frac{h_{01} + h_{10} - (h_{01} - h_{10})^2}{1 - r - s}. \quad (5)$$

Because both $\hat{\Delta}^{(U)}$ and $\hat{\Delta}^{(P)}$ are unbiased estimators of the treatment effect, we can define a

new class of unbiased estimators as $\hat{\Delta}_2 = w\hat{\Delta}^{(P)} + (1-w)\hat{\Delta}^{(U)}$ with $0 \leq w \leq 1$.

The approximate distribution of $\sqrt{n}(\hat{\Delta}_2 - \Delta)$ is normal with mean 0 and variance

$$\text{Var}(\sqrt{n}\hat{\Delta}_2) = w^2\tau_2^2 + (1-w)^2\tau_1^2.$$

The optimal weight, denoted by w^* , which minimizes $Var(\sqrt{n}\hat{\Delta}_2)$, is solved by setting $\partial Var(\sqrt{n}\hat{\Delta}_2)/\partial w = 0$. It is easy to obtain solution $w^* = r_1^2 / (\tau_1^2 + \tau_2^2)$. Thus the new hybrid estimator is

$$\hat{\Delta}_2^* = w^* \hat{\Delta}^{(P)} + (1 - w^*) \hat{\Delta}^{(U)}.$$

It combines inference from two data sources: unpaired outcomes from independent cohorts ($\mathbf{Y}_1, \mathbf{Y}_3$) and paired outcomes from a single cohort ($\mathbf{Y}_2$). It has the smallest variance among all weighted averages of $\hat{\Delta}^{(P)}$ and $\hat{\Delta}^{(U)}$. Under the optimal weight w^* , $\sqrt{n}(\hat{\Delta}_2^* - \Delta)$ is approximately normal with mean 0 and variance

$$\sigma_2^2 = \frac{\tau_1^2 \tau_2^2}{\tau_1^2 + \tau_2^2} = \left(1/\tau_1^2 + 1/\tau_2^2\right)^{-1}.$$

Note that σ_2^2 is the harmonic mean of τ_1^2 and τ_2^2 . It is always true that $\sigma_2^2 \leq \min\{\tau_1^2, \tau_2^2\}$. i.e., the hybrid estimator $\hat{\Delta}_2^*$ is more efficient than the original estimators $\hat{\Delta}^{(P)}$ and $\hat{\Delta}^{(U)}$. Furthermore, by rewriting w^* as

$$w^* = \frac{1/\tau_2^2}{1/\tau_1^2 + 1/\tau_2^2},$$

it is clear that between $\hat{\Delta}^{(P)}$ and $\hat{\Delta}^{(U)}$, the one with the better precision is assigned the larger weight in the hybrid estimator $\hat{\Delta}_2^*$.

Calculating the optimal weight w^* requires the values of (p_0, p_1, h_{11}) . In practice, their true values are usually unknown, and are replaced by observed sample proportions.

2.3. Comparison of $\hat{\Delta}_1^*$ and $\hat{\Delta}_2^*$

Due to the independence between $\mathbf{Y}_1$ and $\mathbf{Y}_2$, as well as between $\mathbf{Y}_2$ and $\mathbf{Y}_3$, the weights in $\hat{\Delta}_1^*$, $\phi_0 = m_2/(m_1 + m_2)$ and $\phi_1 = m_2/(m_2 + m_3)$, are optimal in minimizing the variances of $\hat{p}_0^{(H)}$ and $\hat{p}_1^{(H)}$, respectively. However, when the paired outcomes from the same subject are correlated, minimizing the individual variances of $\hat{p}_0^{(H)}$ and $\hat{p}_1^{(H)}$ does not guarantee minimizing the variance of their difference $\hat{\Delta}_1 = \hat{p}_1^{(H)} - \hat{p}_0^{(H)}$. That is, $\hat{\Delta}_1^*$ might not be optimal among the class of estimators $\hat{\Delta}_1$. The proposed estimator $\hat{\Delta}_2^*$ is advantageous in that it accounts for the pairing of observations in $\mathbf{Y}_2$ and the optimal weight w^* directly minimizes the variance of the treatment effect estimator $\hat{\Delta}_2 = w \hat{\Delta}^{(P)} + (1 - w) \hat{\Delta}^{(U)}$. It should be pointed out, however, that the optimality of $\hat{\Delta}_2^*$ is achieved within the class of estimators $\hat{\Delta}_2$, which is a subclass of $\hat{\Delta}_1$ under the constraint $\phi_0 = \phi_1$. Specifically,

$$\begin{aligned}
 \hat{\Delta}_2 &= w\hat{\Delta}^{(P)} + (1-w)\hat{\Delta}^{(U)} \\
 &= w(\hat{p}_1^{(P)} - \hat{p}_0^{(P)}) + (1-w)(\hat{p}_1^{(U)} - \hat{p}_0^{(U)}) \\
 &= [w\hat{p}_1^{(P)} + (1-w)\hat{p}_1^{(U)}] - [w\hat{p}_0^{(P)} + (1-w)\hat{p}_0^{(U)}].
 \end{aligned}$$

We have the following result concerning the relative efficiency between $\hat{\Delta}_1^*$ and $\hat{\Delta}_2^*$:

Theorem 1—If $r = s$, then $\sigma_2^2 \leq \sigma_1^2$. The “=” sign holds only when $\rho = 0$.

Proof: See Appendix B.

Theorem 1 offers assurance that when the proportions of subjects with missing outcomes are balanced between the two treatments ($r = s$), the proposed hybrid estimator $\hat{\Delta}_2^*$ will always be more efficient than the traditional estimator $\hat{\Delta}_1^*$. There is no definite conclusion where $r \neq s$.

In Figure 1 we plot the relative efficiency between $\hat{\Delta}_1^*$ and $\hat{\Delta}_2^*$, defined as $R = \sigma_1^2 / \sigma_2^2$, against correlation coefficient ρ , under different combinations of $(p_0; p_1; r, s)$. We declare $\hat{\Delta}_2^*$ to be more efficient than $\hat{\Delta}_1^*$, if $R > 1$, and vice versa. In each graph, the title indicates the combination of $(p_0; p_1)$ under investigation. The horizontal axis shows the value of ρ . Note that the valid range of ρ are different depending on $(p_0; p_1)$ in each graph. The vertical axis represents the relative efficiency R . Equal efficiency $R = 1$ is marked by a horizontal dashed line. We explore four missing patterns ($r = 0.1, s = 0.1$), ($r = 0.2, s = 0.2$), ($r = 0.1, s = 0.3$), and ($r = 0.3, s = 0.1$), represented by colors black, red, blue, and, green, respectively. Figure 1 empirially confirms the conclusion of Theorem 1, where the black and red curves (both under $r = s$) are always above the line of equal efficiency. When $r \neq s$, however, we observe that $\hat{\Delta}_1^*$ tends to be more efficient than $\hat{\Delta}_2^*$ if the within-subject correlation is negative or weakly positive. Such scenarios are relatively rare in paired observations from pre-post studies. The proposed estimator $\hat{\Delta}_2^*$ becomes superior when there is a moderate to strong positive correlation. Such scenarios are more frequently the case in practice. We further observe that although $\hat{\Delta}_1^*$ might be more efficient than $\hat{\Delta}_2^*$, it is only by a small margin when it does. Overall all the scenarios we have explored, the smallest relative efficiency we have observed is $R = 0.949$. That is, $\hat{\Delta}_1^*$ is at most 5% more efficient than $\hat{\Delta}_2^*$. On the other hand, the largest relative efficiency we observed is $R = 2.55$, at $(p_0 = 0.4, p_1 = 0.45, r = 0.3, s = 0.1)$ and $\rho = 0.9$. That is, $\hat{\Delta}_2^*$ can be substantially more efficient than $\hat{\Delta}_1^*$ when there is a strong positive correlation between paired outcomes. Finally, with the formulae of σ_1^2 and σ_2^2 provided, it is straightforward for practitioners to numerically examine the relative efficiency of $\hat{\Delta}_1^*$ versus $\hat{\Delta}_2^*$ based on assumed (p_0, p_1, ρ, r, s) at the design stage.

2.4. Hypothesis Testing and Sample Size Determination

Under the proposed estimator $\hat{\Delta}_2^*$, in order to test null hypothesis $H_0 : p_0 - p_1 = 0$, we calculate test statistics

$$Z = \frac{|\hat{\Delta}_2^*|}{\sqrt{\sigma_2^2/n}} = \frac{\sqrt{n}|w^*\hat{\Delta}^{(P)} + (1-w^*)\hat{\Delta}^{(U)}|}{\sqrt{w^{*2}\tau_2^2 + (1-w^*)^2\tau_1^2}}, \quad (6)$$

and reject H_0 if $Z > z_{1-\alpha/2}$. Here $z_{1-\alpha/2}$ is the 100(1 - α)th percentile of the standard normal distribution. Note that based on $(\mathbf{Y}_1, \mathbf{Y}_3)$ only, the test statistic $Z_1 = \hat{\Delta}^{(U)} / \sqrt{\tau_1^2/n}$ is widely used to compare proportions between two independent groups. Based on $\mathbf{Y}_2$ only, the test statistic $Z_2 = \hat{\Delta}^{(P)} / \sqrt{\tau_2^2/n}$ can be shown to be equivalent to the McNemar's test statistic. Equation (6) presents Z as a hybrid test statistic obtained from Z_1 and Z_2 .

As for sample size determination, under the alternative hypothesis $H_a : p_1 - p_0 = \Delta$, achieving a testing power of $1 - \gamma$ implies that

$$P(Z > z_{1-\alpha/2} | H_a) = 1 - \gamma. \quad (7)$$

Here γ is the type II error. From (7) we have a closed-form sample size solution

$$n_2 = \frac{\sigma_2^2 (z_{1-\alpha/2} + z_{1-\gamma})^2}{\Delta^2} = \frac{\tau_1^2 \tau_2^2 (z_{1-\alpha/2} + z_{1-\gamma})^2}{(\tau_1^2 + \tau_2^2) \Delta^2},$$

where τ_1^2 and τ_2^2 are presented in (4) and (5).

Through similar steps, we can derive a close-form sample size formula based on estimator $\hat{\Delta}_1^*$. We denote this sample size by

$$n_1 = \frac{\sigma_1^2 (z_{1-\alpha/2} + z_{1-\gamma})^2}{\Delta^2},$$

where the expression of σ_1^2 is (3).

3. Simulation

We perform simulation studies to assess the performance of the proposed sample size formula n_2 . Combinations of (p_0, p_1, r, s) similar to those in Figure 1 are explored. We investigate four non-negative values of within-subject correlation, $\rho = 0, 0.1, 0.25$, and 0.5 , because in biomedical studies we seldom encounter a negative correlation and the above values all fall within the valid range of ρ under the various pairs of (p_0, p_1) considered. We

set the nominal levels of power at $1 - \gamma = 0.8$ and type I error at $\alpha = 0.05$. For each combination of (p_0, p_1, ρ, r, s) , we first calculate sample size n_2 . Then 10,000 data sets are generated, each contains n_2 pairs of binary outcomes following (p_0, p_1, ρ) , denoted by $\{(Z_{i0}, Z_{i1}), i = 1, \dots, n_2\}$. To impose missing values, we generate a random vector of indicators $\mathbf{g} = (g_1, \dots, g_{n_2})$, where g_i are independent and follow a multinomial distribution, taking values 1, 2, and 3, with probabilities $r, (1 - r - s)$, and s , respectively. We set Z_{i1} to be missing if $g_i = 1$, and set Z_{i0} to be missing if $g_i = 3$. The final simulated dataset is defined by $\mathbf{Y}_1 = \{Z_{i0} : g_i = 1\}$, $\mathbf{Y}_2 = \{(Z_{i0}, Z_{i1}) : g_i = 2\}$, and $\mathbf{Y}_3 = \{Z_{i1} : g_i = 1\}$. Then hypothesis testing based on (6) is conducted on data set. Empirical power and type I error are computed as the proportion of times H_0 is rejected under the alternative and null hypothesis, respectively. A similar simulation study is performed for n_1 .

In Tables 2 and 3 we show the simulation results for n_1 and n_2 . In the last three columns, we present the calculated sample size together with empirical power $1 - \hat{\gamma}$ and empirical type I error $\hat{\alpha}$ under different design configurations. In the column named n_0 , we list the sample sizes based on the McNemar's test [4] where all subject contribute complete pairs of observations ($r = s = 0$). Over the scenarios that we have explored, the calculated sample sizes have a wide range, from 120 to 2308, and the empirical powers and type I errors are generally close to their nominal levels, indicating good performance of the sample size formulas. We also observe that the strength of within-subject correlation has a profound impact on sample size requirement. For example, under marginal probabilities ($p_0 = 0.4, p_1 = 0.45$) and missing pattern ($r = 0.1, s = 0.1$), if we assume no within-subject correlation ($\rho = 0$), the required sample size is 1701. If we assume a correlation of $\rho = 0.5$, the sample size drops to 901, a 47% reduction. Thus at the design stage of experiments with paired outcomes, it is critical to conduct literature review or preliminary analysis to evaluate within-subject correlation. Furthermore, the simulations demonstrate that utilizing incomplete data does lead to saving in sample size. For example, under ($p_0 = 0.25, p_1 = 0.35, \rho = 0.25$), if we exclude incomplete data from analysis, then in order to obtain $n_0 = 251$ complete pairs of outcomes to achieve 80% power at 5% significance level, we need to enroll 314 ($251/0.8$) and 419 ($251/0.6$) unique subjects under missing patterns ($r = 0.1, s = 0.1$) and ($r = 0.2, s = 0.2$), respectively. The proposed sample size n_2 indicates that, if we appropriately incorporate incomplete data into statistical inference, 280 and 327 unique subjects are actually needed to achieve the desired power and type I error, both substantially smaller than those required when ignoring incomplete data. Similar saving in sample size is observed on n_1 . Finally, comparing Tables 2 and 3, under balanced missing patterns, ($r = 0.1, s = 0.1$) and ($r = 0.2, s = 0.2$), the sample sizes (n_2) in Table 2 are always smaller than their counter parts (n_1) in Table 3. Under the unbalanced missing pattern, ($r = 0.1, s = 0.3$), the sample sizes in Table 3 can be smaller. However, when within-subject correlation is moderately strong ($\rho = 0.5$), the sample sizes in Table 2 are all smaller than those in Table 3. This observation empirically confirms the conclusion about relative efficiency in Theorem 1. Due to space limit, we omit the simulation results under missing pattern ($r = 0.3, s = 0.1$).

4. A Real Application Example

Evans *et al.* [17] performed a cohort study on mothers' mood through pregnancy and after childbirth. The Edinburgh postnatal depression scale was implemented on each woman at 18

and 32 weeks of pregnancy, as well as 8 weeks and 8 months postpartum. 13,799 women were identified as eligible, among whom 12,059 completed at least one of the four questionnaires and 9,028 completed all four. The authors performed analysis based on the 9,028 women with complete data. The outcome of primary interest is whether a woman had depressive disorder at the four stages, indicated by a score above 12. It was reported that the numbers (%) of women with depressive disorder are 1062 (11.8%) at 18 weeks of pregnancy, 1222 (13.5%) at 32 weeks of pregnancy, 821 (9.1%) at 8 weeks postpartum, and 738 (8.2%) at 8 months postpartum. The joint probabilities of depression at the four stages were not reported.

Suppose we would like to conduct a similar study to determine whether there is a significant difference in the risk of depression for women between the stages at 18 weeks of pregnancy and 8 weeks postpartum. During these two stages women undergo dramatical changes both physically and psychologically. We assume that the true incidence rates of depression at these two stages are $p_0 = 11.8\%$ and $p_1 = 9.1\%$, respectively. Because in the previous study $9028/12059 = 75\%$ of subjects contributed complete data, we specify $1 - r - s = 75\%$. We assume the incomplete data to be evenly distributed between the two stages, thus $r = s = 12.5\%$. To detect a difference of $\delta = p_1 - p_0 = -2.7\%$ with 90% power and 5% type I error, the sample sizes calculated based on n_2 3077, 2538, 1966, 1667, 1359, and 713, under within-subject correlation $\rho = 0, 0.2, 0.4, 0.5, 0.6$, and 0.8 , respectively. The sample sizes calculated based on n_1 under the same configurations are 3077, 2553, 2029, 1767, 1505, 981.

For illustration purposes, in the above calculation we have assumed MCAR. In studies of mood it is likely that the probabilities of missing data might depend on observed or unobserved data. In such cases an additional model for the missing mechanism is needed to achieve valid inference, and simulation studies are needed to assess sample size requirement.

5. Discussion

Paired binary outcomes with potentially incomplete data are frequently collected and analyzed in biomedical and social behavioral studies. In this paper we propose a new estimator for the treatment effect which appropriately accounts for the mixed nature of the data and directly minimizes variance for the estimator of interest. We theoretically prove that under balanced missing patterns ($r = s$), the proposed estimator is always more efficient than the traditional estimator. We demonstrate through simulation that under unbalanced missing patterns ($r \neq s$) the proposed estimator will be superior when there is a moderate to strong positive correlation between paired outcomes. We also derive close-form sample size formulas to help researchers determine how many unique subjects need to be enrolled in this type of studies. The sample size formulas are shown to preserve the nominal levels of power and type I error under various design configurations.

One limitation of the proposed estimator is that $\hat{\Delta}_2$ is not well defined when outcomes are observed from all subjects under one of the treatments ($r = 0$ or $s = 0$), which is not the case for the traditional estimator $\hat{\Delta}_1$. Both $\hat{\Delta}_1$ and $\hat{\Delta}_2$ are developed based on normal approximation under the large sample theory. Thus the performance of the estimators and

the corresponding sample size formulas depends on the quality of normal approximation. It is widely known that the normal approximation is less satisfactory under extreme values of probabilities (p_0 and p_1 close to 0 or 1) or small sample sizes. Here we need to check the numbers of observations for both paired and unpaired outcomes. When normal approximation is unsatisfactory, nonparametric exact methods should be implemented, which will be one of our future research topics.

The proposed sample size formula is derived based on the MCAR assumption. When the MCAR assumption does not hold, an additional model accounting for the missing mechanism is required to achieve valid inference. Because the missing mechanism is usually unknown and unique in each study, it is unlikely to derive a universally applicable sample size formula for studies without the MCAR assumption. In such cases, sample size requirement should be evaluated using customized simulation schemes designed for particular studies, which is beyond the scope of this paper. It is generally expected, however, that a more complicated model requires a larger sample size to maintain desired power and type I error.

Acknowledgement

The work was supported in part by NIH grants 1UL1TR001105, AHRQ grant R24HS22418, CPRIT grants RP110562-C1 and RP120670-C1, and NSF grant IIS-1302497-03.

Appendix A. The Derivation of σ^2

$$\begin{aligned} Var(\hat{\Delta}_1) &= Var\left(\frac{\sum_{i=1}^{m_1+m_2} Y_{i0}}{m_1+m_2}\right) + Var\left(\frac{\sum_{i=m_1+1}^n Y_{i1}}{m_2+m_3}\right) - 2Cov\left(\frac{\sum_{i=1}^{m_1+m_2} Y_{i0}}{m_1+m_2}, \frac{\sum_{i=m_1+1}^n Y_{i1}}{m_2+m_3}\right) \\ &= \frac{p_0(1-p_0)}{m_1+m_2} + \frac{p_1(1-p_1)}{m_2+m_3} - \frac{Cov\left(\sum_{i=1}^{m_1+m_2} Y_{i0}, \sum_{i=m_1+1}^n Y_{i1}\right)}{(m_1+m_2)(m_2+m_3)}. \end{aligned}$$

Here we use the fact that $\mathbf{Y}_1$ and $\mathbf{Y}_3$ are independent. Because

$$\begin{aligned} Cov\left(\sum_{i=m_1+1}^{m_1+m_2} Y_{i0}, \sum_{i=m_1+1}^{m_1+m_2} Y_{i1}\right) &= \sum_{i=m_1+1}^{m_1+m_2} \sum_{j=m_1+1}^{m_1+m_2} Cov(Y_{i0}, Y_{j1}) \\ &= m_2 \left[Cov(Y_{i0}, Y_{i1}) + \sum_{j \neq i} Cov(Y_{i0}, Y_{j1}) \right] \\ &= m_2 Cov(Y_{i0}, Y_{i1}) \\ &= m_2 [E(Y_{i0}, Y_{i1}) - E(Y_{i0}) E(Y_{i1})] = m_2 (h_{11} - p_0 p_1) \quad (8) \end{aligned}$$

Plug (8) into $Var(\hat{\Delta}_1)$ and we obtain the expression of $\sigma^2(H^\dagger)$.

Appendix B. Proof of Theorem 1

We define $e_t^2 = Var(Y_{it}) = p_t(1-p_t)$ for $t = 0, 1$. It can also be shown that

$$Var(d_i) = h_{01} + h_{10} - (h_{01} - h_{10})^2 = Var(Y_{i0}) + Var(Y_{i1}) - 2Cov(Y_{i0}, Y_{i1}) = e_0^2 + e_1^2 - 2\rho e_0 e_1.$$

Thus under $r = s$, we can simplify the formulas of σ_1^2 and σ_2^2 as

$$\begin{aligned}\sigma_1^2 &= \frac{e_0^2 + e_1^2}{1-r} - \frac{2(1-2r)\rho e_0 e_1}{(1-r)^2}, \\ \sigma_2^2 &= \left(\frac{r}{e_0^2 + e_1^2} + \frac{1-2r}{e_0^2 + e_1^2 - 2\rho e_0 e_1} \right)^{-1}.\end{aligned}$$

After a few algebra steps, we can show that

$$\sigma_1^2 - \sigma_2^2 = \frac{4r(1-2r)\rho^2 e_0^2 e_1^2}{(1-r)^2 [(1-2r)(e_0^2 + e_1^2) + r(e_0^2 + e_1^2 - 2\rho e_0 e_1)]}.$$

Because $Var(d_i) = e_0^2 + e_1^2 - 2\rho e_0 e_1 \geq 0$ and $(1-2r) > 0$ is the proportion of subjects with complete pairs of observations, we have

$$\sigma_1^2 - \sigma_2^2 \geq 0$$

and the “=” holds only when $\rho = 0$. Thus we complete the proof.

References

1. Poletto FG, Singer JM, Paulino CD. Comparing diagnostic tests with missing data. *Journal of Applied Statistics*. 2011; 38(6):1207–1222.
2. Campbell G, Pennello G, Yue L. Missing data in the regulation of medical devices. *American Journal of Public Health*. 2011; 21(2):180–195.
3. Miettinen OS. The matched pairs design in the case of all-or-none responses. *Biometrics*. 1968; 24(2):339–352. [PubMed: 5683874]
4. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*. 1987; 43(1):207–211. [PubMed: 3567305]
5. Guenther WC. Power and Sample Size for Approximate Chi-Square Tests. *The American Statistician*. 1977; 3(2):83–5.
6. Ekbohm G. On testing the equality of proportions in the paired case with incomplete data. *Psychometrika*. 1982; 47(1):115–118.
7. Choi SC, Stablein DM. Practical tests for comparing two proportions with incomplete data. *Applied Statistics*. 1982; 31:256–62.
8. Thomson PC. A hybrid paired and unpaired analysis for the comparison of proportions. *Statistics in medicine*. 1995; 14(13):1463–70. [PubMed: 7481184]
9. Shih WJ. Maximum likelihood estimation and likelihood ratio test with incomplete pairs. *Journal of Statistical Computation and Simulation*. 1987; 21:187–94.
10. Choi SC, Stablein DM. Comparing incomplete paired binomial data under non-random mechanisms. *Statistics in Medicine*. 1988; 7:929–39. [PubMed: 3175392]
11. Tang M, Tang N. Exact tests for comparing two paired proportions with incomplete data. *Biometrical Journal*. 2004; 46(1):72–82.

12. Tang M, Ling M, Tian G. Exact and approximate unconditional confidence intervals for proportion difference in the presence of incomplete data. *Statistics in medicine*. 2009; 28(4):625–41. [PubMed: 19035467]
13. LI HQ, Tian GL, Jiang XJ, Tang NS. Testing hypothesis for a simple ordering in incomplete contingency tables. *Computational Statistics & Data Analysis*. 2016; 99:25–37.
14. LI HQ, Chan ISF, Tang ML, Tian GL, Tang NS. Confidence interval construction for rate ratio in matched-pair studies with incomplete data. *Journal of Biopharmaceutical Statistics*. 2014; 24(3): 546–568. [PubMed: 24697611]
15. Chaganty NR, Joe H. Range of correlation matrices for dependent Bernoulli random variables. *Biometrika*. 2006; 93(1):197–206.
16. Tang NS, LI HQ, Tang ML, LI J. Confidence interval construction for the difference between two correlated proportions with missing observations. *Journal of Biopharmaceutical Statistics*. 2016; 26(2):323–338. [PubMed: 25632882]
17. Evans J, Heron J, Francomb H, Oke S, Golding J. Cohort study of depressed mood during pregnancy and after childbirth. *British medical journal*. 2001; 323(7307):257–60. [PubMed: 11485953]

Figure 1.

The relative efficiency between $\hat{\Delta}_1$ and $\hat{\Delta}_2$, defined as $R = \sigma_1^2 / \sigma_2^2$, versus correlation ρ under various combinations of (p_0, p_1) . The horizontal axis shows ρ and the vertical axis shows R . The horizontal dashed line marks equal efficiency $R = 1$. Four missing patterns are explored: $(r = 0.1, s = 0.1)$, $(r = 0.2, s = 0.2)$, $(r = 0.1, s = 0.3)$, and $(r = 0.3, s = 0.1)$, represented by colors black, red, blue and green, respectively.

Table 1

Probabilities of Pre- and Post-Intervention Outcomes

Pre-intervention	Post-intervention		
	NO	YES	
NO	h_{00}	h_{01}	$1 - p_0$
YES	h_{10}	h_{11}	p_0
		$1 - p_1$	p_1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2The performance of sample sizes n_2

(p_0, p_1)	ρ	n_0	$(r = 0.1, s = 0.1)$	$(r = 0.2, s = 0.2)$	$(r = 0.1, s = 0.3)$
(0.10, 0.15)	0	689	759(0.801, 0.050)	854(0.799, 0.048)	894(0.810, 0.055)
	0.10	622	692(0.806, 0.052)	789(0.806, 0.055)	824(0.813, 0.052)
	0.25	521	588(0.809, 0.046)	686(0.809, 0.051)	712(0.818, 0.057)
	0.50	353	408(0.811, 0.053)	494(0.809, 0.050)	508(0.824, 0.060)
(0.10, 0.20)	0	202	219(0.797, 0.052)	246(0.798, 0.052)	254(0.826, 0.078)
	0.10	183	200(0.803, 0.049)	228(0.806, 0.052)	235(0.830, 0.083)
	0.25	155	171(0.799, 0.054)	199(0.809, 0.052)	204(0.827, 0.084)
	0.50	108	120(0.812, 0.046)	145(0.822, 0.051)	148(0.848, 0.073)
(0.25, 0.30)	0	1254	1387(0.795, 0.055)	1560(0.797, 0.048)	1655(0.806, 0.053)
	0.10	1129	1263(0.804, 0.054)	1441(0.803, 0.054)	1521(0.807, 0.052)
	0.25	942	1071(0.804, 0.050)	1249(0.799, 0.047)	1309(0.806, 0.052)
	0.50	631	736(0.803, 0.052)	893(0.796, 0.052)	923(0.807, 0.051)
(0.25, 0.35)	0	332	362(0.802, 0.056)	408(0.799, 0.053)	430(0.799, 0.052)
	0.10	299	330(0.798, 0.051)	376(0.797, 0.051)	396(0.803, 0.053)
	0.25	251	280(0.813, 0.048)	327(0.809, 0.054)	341(0.810, 0.053)
	0.50	170	193(0.803, 0.050)	234(0.804, 0.054)	241(0.811, 0.053)
(0.40, 0.45)	0	1537	1701(0.794, 0.053)	1914(0.795, 0.044)	2038(0.798, 0.053)
	0.10	1383	1548(0.805, 0.050)	1767(0.798, 0.046)	1872(0.812, 0.048)
	0.25	1154	1312(0.806, 0.047)	1531(0.798, 0.046)	1610(0.793, 0.052)
	0.50	771	901(0.803, 0.049)	1094(0.800, 0.049)	1133(0.802, 0.052)
(0.40, 0.50)	0	391	428(0.803, 0.052)	481(0.805, 0.052)	512(0.802, 0.049)
	0.10	353	389(0.796, 0.051)	444(0.801, 0.049)	471(0.801, 0.053)
	0.25	294	330(0.803, 0.051)	385(0.797, 0.046)	405(0.801, 0.051)
	0.50	198	227(0.802, 0.053)	275(0.798, 0.052)	285(0.805, 0.052)

Simulation results based on nominal power 80% and type I error 5%. In each cell we show $n_2(1 - \hat{\gamma}, \hat{\alpha})$, the calculated sample size, empirical power, and empirical type I error. The column named n_0 lists sample sizes based on the McNemar's test where all subjects contribute complete pairs of outcomes ($r = s = 0$).

Table 3The performance of sample sizes: n_1

(p_0, p_1)	ρ	n_0	$(r = 0.1, s = 0.1)$	$(r = 0.2, s = 0.2)$	$(r = 0.1, s = 0.3)$
(0.10, 0.15)	0	689	759(0.801, 0.050)	854(0.799, 0.048)	849(0.807, 0.051)
	0.10	622	693(0.803, 0.050)	791(0.810, 0.052)	785(0.791, 0.058)
	0.25	521	593(0.810, 0.049)	696(0.805, 0.051)	689(0.800, 0.047)
	0.50	353	427(0.805, 0.051)	539(0.800, 0.050)	529(0.805, 0.050)
(0.10, 0.20)	0	202	219(0.808, 0.052)	246(0.803, 0.048)	241(0.794, 0.053)
	0.10	183	200(0.803, 0.049)	228(0.804, 0.051)	223(0.803, 0.049)
	0.25	155	172(0.818, 0.052)	202(0.819, 0.052)	196(0.800, 0.051)
	0.50	108	126(0.820, 0.049)	157(0.817, 0.049)	151(0.807, 0.052)
(0.25, 0.30)	0	1254	1387(0.803, 0.051)	1560(0.804, 0.051)	1574(0.807, 0.051)
	0.10	1129	1264(0.798, 0.047)	1444(0.800, 0.049)	1455(0.794, 0.050)
	0.25	942	1079(0.801, 0.047)	1268(0.796, 0.055)	1277(0.796, 0.048)
	0.50	631	772(0.800, 0.044)	976(0.800, 0.054)	981(0.805, 0.050)
(0.25, 0.35)	0	332	362(0.806, 0.051)	408(0.801, 0.049)	409(0.796, 0.053)
	0.10	299	330(0.803, 0.050)	377(0.796, 0.058)	378(0.798, 0.052)
	0.25	251	282(0.806, 0.051)	332(0.803, 0.053)	332(0.806, 0.052)
	0.50	170	202(0.804, 0.054)	256(0.799, 0.052)	255(0.800, 0.056)
(0.40, 0.45)	0	1537	1701(0.801, 0.049)	1914(0.801, 0.053)	1940(0.801, 0.050)
	0.10	1383	1550(0.800, 0.050)	1770(0.802, 0.048)	1795(0.802, 0.055)
	0.25	1154	1323(0.803, 0.050)	1555(0.800, 0.047)	1576(0.802, 0.050)
	0.50	771	945(0.803, 0.053)	1196(0.797, 0.053)	1212(0.796, 0.049)
(0.40, 0.50)	0	391	428(0.798, 0.049)	481(0.796, 0.051)	488(0.798, 0.054)
	0.10	352	390(0.805, 0.055)	445(0.802, 0.053)	451(0.800, 0.054)
	0.25	294	333(0.806, 0.054)	391(0.802, 0.049)	396(0.800, 0.051)
	0.50	198	238(0.802, 0.049)	301(0.799, 0.052)	305(0.803, 0.051)

Simulation results based on nominal power 80% and type I error 5%. In each cell we show $n_2(1 - \hat{\gamma}, \hat{\alpha})$, the calculated sample size, empirical power, and empirical type I error. The column named n_0 lists sample sizes based on the McNemar's test where all subjects contribute complete pairs of outcomes ($r = s = 0$).