Statistics
in Medicine

Research Article

Received 22 May 2012, Accepted 24 May 2013 Published online 9 July 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5887

A hybrid method in combining treatment effects from matched and unmatched studies

Jinyoung Byun,[a,c] Dejian Lai,[a,b*†] Sheng Luo,[a] Jan Risser,[a] Betty Tung[a] and Robert J. Hardy[a]

The most common data structures in the biomedical studies have been matched or unmatched designs. Data structures resulting from a hybrid of the two may create challenges for statistical inferences. The question may arise whether to use parametric or nonparametric methods on the hybrid data structure. The Early Treatment for Retinopathy of Prematurity study was a multicenter clinical trial sponsored by the National Eye Institute. The design produced data requiring a statistical method of a hybrid nature. An infant in this multicenter randomized clinical trial had high-risk prethreshold retinopathy of prematurity that was eligible for treatment in one or both eyes at entry into the trial. During follow-up, recognition visual acuity was accessed for both eyes. Data from both eyes (matched) and from only one eye (unmatched) were eligible to be used in the trial. The new hybrid nonparametric method is a meta-analysis based on combining the Hodges–Lehmann estimates of treatment effects from the Wilcoxon signed rank and rank sum tests. To compare the new method, we used the classic meta-analysis with the t-test method to combine estimates of treatment effects from the paired and two sample t-tests. We used simulations to calculate the empirical size and power of the test statistics, as well as the bias, mean square and confidence interval width of the corresponding estimators. The proposed method provides an effective tool to evaluate data from clinical trials and similar comparative studies. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: hybrid design; matched and unmatched studies; meta-analysis; nonparametric methods; Wilcoxon rank sum; Wilcoxon signed rank

1. Introduction

In many biomedical studies, study subjects are usually paired for a more effective comparison of the treatment. For example, one may want to verify an interventional treatment effect on treating eye diseases. When both eyes in a patient have disease and are available for randomization, it may be feasible to randomly assign one eye of the patient to the treatment or control. Then, the other eye is automatically allocated to the other group. Hence, we have a matched design. In some cases, the patient may only have one eye eligible for treatment. Therefore, only the eligible diseased eye would be randomly allocated to either treatment or control, and we have an unmatched study [1–5]. This situation is common to studies on eyes, ears, knees, arms and kidneys, and the outcome leads to clustered measurements.

To obtain an overall treatment effect assessment, it is necessary to combine the matched and unmatched data. Statistical inferences on overall effectiveness are based on combining results from individuals to reduce random fluctuations. Randomized clinical trials with enough sample sizes have become an integrated part of drug discovery and evaluation. In many instances, meta-analysis is used to combine several studies to examine the overall treatment effect [6] .

[a] *University of Texas School of Public Health, Houston, TX 77030, U.S.A.*
[b] *Faculty of Statistics, Jiangxi University of Finance and Economics, Nanchang, China*
[c] *Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, U.S.A.*
**Correspondence to: Dejian Lai, University of Texas School of Public Health 1200 Herman Pressler, RAS 1006, Houston, TX 77030, U.S.A.*
[†] *E-mail: Dejian.lai@uth.tmc.edu*

Ophthalmologists are confronted with infants that have bilateral disease (two eyes eligible for treatment) and unilateral disease (one eye eligible). The challenge that we address in this manuscript is the analysis of the full range of visual acuity data at age 6 on an ordinal scale where some patients have good vision and others may have more severe impairment in their measurable visual acuity data. The remainder of the participants may have little or no measurable acuity. The information in some of these clinical trials resides in both matched and unmatched portions of the design. The effect estimator and test statistics that can extract and combine the information from both portions is usually preferable.

It is crucial that both matched and unmatched pairs be properly combined in evaluating the treatment effect to gain as much information as possible for statistical and clinical inferences for the patient population. Many methods of combining odd ratios, relative risks and mean differences across studies are available in the literature [6]. Wilcoxon signed rank test and Wilcoxon rank sum test are the two most commonly used nonparametric test statistics applied to matched and unmatched studies, respectively. However, there seemed a lack of methods of combining results from these tests.

In this article, we propose a hybrid procedure of combining the classic Wilcoxon signed rank test and the Wilcoxon rank sum test to assess the combined treatment effects from matched and unmatched studies. The proposed hybrid procedure would properly evaluate the treatment effects of the therapies in improving and maintaining vision and quality of life and reducing health care costs. We illustrate our proposed statistical procedure through data analysis of outcome measurements collected from Early Treatment for Retinopathy of Prematurity (ETROP) study. The proposed method would also be useful for effectively analyzing important outcome data from other studies with a matched and unmatched data structure.

In conventional situations, if the outcome data were continuous and normally distributed, we would use a classic paired t-test and a two sample t-test for the matched and unmatched groups, respectively. The combined effect of the treatment can be weighted inversely to its variance according to well-established tools used in meta-analysis [6].

For normally distributed measurements, researchers can also use linear models with various correlation structures to jointly estimate the treatment effect [7]. For example, let y_{mti} and y_{mci} be the outcome of treated and controlled eyes for matched pair i, respectively. Let the outcome from the jth treated eye be y_{utj} and the outcome from the lth controlled eye be y_{ucl} for the unmatched eyes. Assume that there are n_1 matched pairs, n_2 unmatched treated eyes and n_3 unmatched controlled eyes. We may construct a vector of outcome measures $Y = (y_{mt1}, y_{mc1}, \cdots, y_{mtn_1}, y_{mcn_1}, y_{ut1}, \cdots, y_{utn_2}, y_{uc1}, \cdots, y_{ucn_3})'$.

Let the design matrix $X = (X_1', \cdots, X_{n_1}' X_{ut}', X_{uc}')'$,where $X_i = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, $X_{ut} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$, $X_{uc} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}$. Then, the linear model combining the outcomes from the matched and unmatched is $Y = XB + E$, where $B=(\alpha, \beta)'$ and E is the vector of errors with variance–covariance matrix $\Sigma = \sigma^2 \text{diag}(\Sigma_1, \cdots, \Sigma_{n_1}, 1, \cdots, 1)$, $\Sigma_i = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where ρ is the correlation coefficient of the outcome between the treated and controlled eyes. One can estimate the parameters in the model via various algorithms.

In some epidemiological studies and clinical trials, the outcomes may only be measured by dichotomous values such as 0 and 1. One can estimate odds ratio for matched and unmatched studies through analysis of contingency tables or classic logistic regression modeling. There are various tools proposed for combining treatment effects from matched and unmatched [1, 8].

In many scientific investigations, the outcome measures may be neither normally distributed nor dichotomously valued, for example, in quantifying vision acuity we discussed in Section 2. In Section 3, we introduce our test statistic based on a new way of estimating the shift of location of the outcomes. Following the data structure of an actual clinical trial, we conducted a simulation study on the proposed test statistic with comparisons to the classic t-test statistic in meta-analysis of combining two studies. We present the results of bias, mean square errors and confidence width as well as the empirical type 1 error rates and power in Section 4. The data analysis of an actual clinical trial in ophthalmology is given in Section 5. Some concluding remarks are in Section 6.

2. The early treatment for retinopathy of prematurity study

The ETROP study is a multicenter clinical trial funded by the National Eye Institute. The ETROP study started enrollment on October 1, 2000. Infants were selected for entry into the ETROP clinical trial by using risk model from an earlier trial. A risk of unfavorable outcomes was based on a logistic risk model (RM-ROP2) developed from Cryotherapy for Retinopathy of Prematurity study data [9]. This risk model was used to select eyes at high risk of blindness for entry into the ETROP study. At randomization, one group of eyes was treated with laser or cryotherapy if needed. The other group was treated only if the eye progressed to a point in the disease known as the previously shown 'threshold' for treatment.

In 2003, the ETROP Study Group published the 9-month outcome results for grating acuity that uses Teller acuity cards. The acuity for infants on early treatment of high-risk ROP was compared with infants receiving the standard conventional management (control) [10]. The full array of vision outcomes was part of the study outcome. However, in this phase of the study, the objective was to preserve vision to better than 3.70 cycles per degree (cpd), and vision of 3.70 cpd or worse/lower was declared 'unfavorable'. At randomization, one group of eyes was treated with laser or cryotherapy if needed. The purpose was to see if earlier treatment is actually better than treatment at the conventional threshold at preventing very poor vision (3.70 cpd or worse/lower). The outcome based on grating acuity in the first phase of the study showed a clear benefit in favor of earlier treatment and that the greatest treatment effect was confined to eyes designated as Type 1 and the remainders were designated as Type 2. Type 1 ROP became the clinical approach for the treatment of ROP shortly after 2003.

Phase II of the ETROP study continued to age 6 to obtain visual acuity. At age 6, the visual acuity was measured according to the sizes and directions of letters or other symbols as in Early Treatment Diabetic Retinopathy Study (ETDRS) [10]. A visual acuity value of 20/20 indicates a normal vision, and 20/x indicates that a person with diseased eye can see detail in 20 feet away the same as a person with normal vision can see in x feet away. The denominator of the value, x, resulting from a diseased eye is usually greater than 20. In ETROP study, the denominator was recorded as the ETDRS outcome of the eye, and an artificial large value such as 9999 in the denominator was assigned if the eye was blind. Hence, the denominator of the acuity scale can be used to rank the eye sight, but it does not allow for a metric to measure vision. This is particularly true if one must include categories where the participant only has light perception or is completely blind. Visual acuity measurements are far from normally distributed, and there are many potential vision outcomes in the extremes such as light perception only or blind. Therefore, methods based on t-tests and linear models are not applicable. Of course, with loss of efficiency, one may convert the values into dichotomous outcomes in testing the hypothesis on treatment effects [11].Taking into account the non-normality and the intrinsic relationship of order, we considered nonparametric techniques. Some of the commonly used nonparametric tools are not directly applicable in combining treatment effects from matched and unmatched studies because of different scales in constructing the test statistics.

In addition to possible adverse effects from new therapeutic invention to patients, clinical trials are usually expensive and time consuming to conduct. For example, the ETROP study involved with long-term (6 years) follow-up of 401 children born prematurely at 26 clinical centers. The primary outcome of the ETROP clinical trial is visual acuity. In ETROP study, 31 children died before the 6-year examination, and 28 lost to follow up. The remaining 342 children were followed at age 6. In ETROP, the diseased eyes were classified by two types. The type I eyes were eyes having zone 1 ROP with plus disease, or zone 1 stage 3 without plus disease, or zone 2, stage 2 or 3 ROP with plus disease. Type II eyes were eyes having zone 1, stage 1 or 2 ROP without plus disease or with zone 2, stage 3 ROP without plus disease. Excluding patients without successful ETDRS measurements and infants with discordant types of diseases, for type I eyes, there were 177 infants in the matched pairs. For the unmatched eyes, there were 25 eyes in the treated group and 26 eyes in the conventional management group. For type II eyes, there were 64 in matched pairs. For the unmatched eyes, 8 were in the treatment group, and 3 were in the conventional management group. In this dataset, there were 31 of blind eyes in the matched group and 1 in the unmatched group. The blind eyes were coded as 20/9999. In summary, there were 241 pairs of matched eyes. For the unmatched eyes, there were 33 eyes in treatment group and 29 in the control group. The data structure in ETROP led us to select $n_1 = 240$ and $n_2 = n_3 = 35$ in our simulation study.

3. The statistical procedures

In this section, we review the two commonly used nonparametric methods. For the matched observations, one may use the Wilcoxon signed rank test in evaluating the treatment effect [12]. For unmatched data, one can use the Wilcoxon rank sum test [12]. The associated location estimates (treatment effects) of the Wilcoxon signed rank test and the Wilcoxon rank sum test are based on the Hodges–Lehmann estimates [13].We then combine the Hodges–Lehmann estimates based on the Wilcoxon signed rank test and the Wilcoxon rank sum test to form our hybrid test statistic.

3.1. Wilcoxon signed rank test

Let y_i and x_i be the outcome measures of treated eye and controlled eye from the matched pair i, respectively. Let $z_i = y_i - x_i$; the Wilcoxon signed rank test is defined as

$$T_1 = \sum_{i=1}^{n_1} R_i \psi_i \tag{1}$$

where R_i denotes the rank of $|z_1|,\ldots,|z_{n_1}|$ and ψ_i=1, if $z_i > 0$ and $\psi_i = 0$, if $z_i < 0$ [12]. One can estimate the treatment effect associated with the Wilcoxon signed rank test by $\widehat{\theta_1}$ = median $\left\{{}^{(z_i + z_j)}/_2, i \leqslant j = 1, ..., n_1\right\}$, where ${}^{(z_i + z_j)}/_2$ is called a Walsh average [14]. The median of these Walsh averages is the Hodges–Lehmann estimate for the Wilcoxon signed rank test.

3.2. Wilcoxon rank sum test

If the patient only has one diseased eye, then only one eye will be randomized to the treatment or control group. Hence, we have an unmatched study. The classic Wilcoxon rank sum is applicable in this situation, especially if the observations are not normally distributed. Let y_j and x_l be the outcome measures from the treated eye and the controlled eye, respectively, where $j = 1, \ldots, n_2$ and $l = 1, \ldots, n_3$. Wilcoxon rank sum statistic is a classic test in evaluating treatment effect for two independent samples [12]. It is defined as

$$T_2 = \sum_{j=1}^{n_2} R_j, \tag{2}$$

where R_j denotes the rank of $y_j$among the combined sample of y_j and x_l. To estimate the treatment effect, one can use the median of the cross differences

$$\widehat{\theta_2} = \text{median}\left\{y_j - x_l\right\}, j = 1, ..., n_2; l = 1, ..., n_3.$$

The median of these cross difference is the Hodges–Lehmann estimate based on the Wilcoxon rank sum test.

3.3. The hybrid test statistic

We recognize that Wilcoxon signed rank test and Wilcoxon rank sum test are not in the same scale. It seems no obvious way of combining this two test statistics directly. We propose to construct the test statistic from the estimates of the treatment (median) effects defined previously. The medians are in the same scale so that we can follow the commonly used tool in meta-analysis [15] and form the hybrid test statistic as

$$T = \frac{\frac{1}{\text{var}\left(\widehat{\theta_1}\right)}\widehat{\theta_1} + \frac{1}{\text{var}\left(\widehat{\theta_2}\right)}\widehat{\theta_2}}{\frac{1}{\text{var}\left(\widehat{\theta_1}\right)} + \frac{1}{\text{var}\left(\widehat{\theta_2}\right)}} \tag{3}$$

Various ways of estimating the variance of the estimated median effects are available for independent and identically observed random variables [16, 17]. For example, for independent samples $x_1, ..., x_n$, McKean and Schrader [16] proposed the following as the estimate of var $\left(\widehat{\theta}\right)$:

$$\widehat{\text{var}\left(\widehat{\theta}\right)} = \left\{(x_{(n-c+1)} - x_{(c)})/2(1.96)\right\}^2,$$

where $c = (n+1)/2 - 1.96(n/4)^{1/2}$ and it is rounded to the near nonzero integer. It was shown that $\widehat{\text{var}\left(\widehat{\theta}\right)}$ is a consistent estimator [17]. However, in our application, for Wilcoxon signed rank test and Wilcoxon rank sum test, the observations are not from independent samples. That is, the Walsh averages $(z_i + z_j)/2$ in Wilcoxon signed rank test are not independent. So are not the cross differences $y_j - x_l$ in Wilcoxon rank sum test. In the next section, we discuss the methods for estimating the variance of the Hodges–Lehmann estimates. The variance estimates were shown to be consistent [18].

3.4. *Variance estimates*

Let $w_{(k)}$ be the kth order statistic of the Walsh average $(w_{(1)} \leqslant \cdots \leqslant w_{(M)})$, where $M = n_1(n_1+1)/2$ is the total number of Walsh averages. Then, a (1-α) symmetric two-sided confidence interval for the treatment effect θ is

$$\left(w_{(c)}, w_{(M+1-c)}\right)$$

where $c = M + 1 - t_{\alpha/2}$ and $t_{\alpha/2}$ is the upper $(\alpha/2)$th percentile the null distribution of the Wilcoxon signed rank statistic. Assuming asymptotic normality [9] for $\widehat{\theta_1}$ and letting $w_{(c)} = \widehat{\theta_1} - Z_{1-\alpha/2}\sigma_{\widehat{\theta_1}}$ and $w_{(M+1-c)} = \widehat{\theta_1} + Z_{1-\alpha/2}\sigma_{\widehat{\theta_1}}$, we obtain an estimate of the variance of $\widehat{\theta_1}$ as

$$\widehat{\text{var}\left(\widehat{\theta_1}\right)} = \left\{(w_{(M+1-c)} - w_{(c)})/2Z_{(1-\alpha/2)}\right\}^2.$$

For an unmatched study, one can construct a (1-α) symmetric two-sided confidence interval for the treatment effect θ for an unmatched study as

$$\left(U_{(c)}, U_{(n_2n_3+1-c)}\right)$$

where $c = n_2n_3 + 1 - t_{\alpha/2}$ and $t_{\alpha/2}$ is the upper $(\alpha/2)$th percentile the null distribution of the Wilcoxon rank sum statistic. Again, assuming asymptotic normality for $\widehat{\theta_2}$ and letting $U_{(c)} = \widehat{\theta_2} - Z_{1-\alpha/2}\sigma_{\widehat{\theta_2}}$ and $U_{(n_2n_3+1-c)} = \widehat{\theta_2} + Z_{1-\alpha/2}\sigma_{\widehat{\theta_2}}$, we obtain an estimate of the variance of $\widehat{\theta_2}$ as

$$\widehat{\text{var}\left(\widehat{\theta_2}\right)} = \left\{\left(U_{(n_2n_3+1-c)} - U_{(c)}\right)/2Z_{(1-\alpha/2)}\right\}^2.$$

These variance estimates are consistent [16].

3.5. *Classic meta-analysis t-test statistic*

Define the classic meta-analysis t-test statistic as

$$T^* = \frac{\frac{1}{\text{var}\left(\widehat{\mu_1}\right)}\widehat{\mu_1} + \frac{1}{\text{var}\left(\widehat{\mu_2}\right)}\widehat{\mu_2}}{\frac{1}{\text{var}\left(\widehat{\mu_1}\right)} + \frac{1}{\text{var}\left(\widehat{\mu_2}\right)}} \tag{4}$$

where $\widehat{\mu_1}$ and $\widehat{\mu_2}$ are the means of the matched and unmatched sample, respectively.

The classic meta-analysis t-test statistic is widely used when both samples are close to realizations from normally distributed random variables. However, it would fail in many non-normal cases including outliers. In the next section, we investigate the performance of the hybrid test statistic and the classic meta-analysis t-test statistic under different settings through simulation.

4. Simulation studies

In Monte Carlo simulations, we generated continuous and discrete observations for matched pairs and unmatched groups for varying sample sizes under the null and several alternatives hypotheses. In this section, we report the empirical size and power of the test statistics from simulation studies to compare the hybrid test statistic with the classic meta-analysis t-test statistic. We also simulated bias, mean square error and confidence interval for the estimates. We simulated the data according to the data structure of the ETROP study and other general data structures. We define the empirical size as the percentages of rejections of the null hypothesis based on the methods applied to the data that are generated under the null hypothesis. We define the empirical power as the percentage of rejections of the null hypothesis based on the methods applied to the data that are generated under various alternatives. We calculated the simulated bias as the average of the differences of the observed value and the expected value, and we defined the simulated mean square error as $\sqrt{\frac{1}{n}\sum(observed-expected)^2}$, where $n = 1000$. We skipped the bias and mean square error calculation for multinomial case under the alternative because the expected value was not available. We calculated the simulated confidence interval width on the basis of large sample approximation as 2 times 1.96 multiplying the standard deviation.

In our simulations, we simulated two components based on the ETROP study. For the matched component, we simulated $n_1 = 240$ pairs of observations, and for unmatched component, we simulated two independent realizations of $n_2 = n_3 = 35$ for treatment and conventional groups, respectively.

4.1. Continuous case

4.1.1. Normal distribution. For the matched study under the null hypothesis, we generated random realizations from bivariate normal distribution with mean 0 and variance 1 with different correlation coefficients (ρ). For the unmatched study under the null hypothesis, we simulated from univariate normal distribution with mean 0 and variance 1. Under the alternative hypotheses, we simulated the empirical statistical power with difference of effect means. For treatment and conventional management groups, we assigned mean 1.1 and 1 with equal variance 1, respectively.

With 1000 replicates, the empirical sizes of the hybrid test statistics were close to the classic meta-analysis t-test statistic. According to the simulations, the empirical statistic powers were similar for both tests. The bias, mean square error and confidence interval width derived from the estimates associated with our hybrid statistic and the meta-analysis t-test were similar. We tabulate the results in Table I. In Table I, we could see that the empirical power increases as the correlation coefficient increases. Our simulation indicated that the classical meta-analysis t-test and the hybrid test statistics have similar sizes and powers under the ideal normal setting. However, for many applications, the observations are far from normal realizations and may counter outliers. In our current simulations, we used the critical values based on large sample properties for both tests that might lead to small empirical type I error than the nominal level in some cases. Because we used the same asymptotic critical value, the sizes of the significant levels were under control, and it is meaningful to compare the empirical powers. To further investigate these issues, we also performed simulation studies on observations resulted from log-normal, multinomial and outliers.

4.1.2. Log-normal distribution. The second simulation study of continuous cases was based on the log-normal distribution. We simulated the observations from log-normal random variable by $\mathrm{Y} = \exp(\mathrm{X})$, where X is an observation from a bivariate normal random variable with mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and variance–covariance matrix $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The exponential transformation was performed for each component. The realizations of Y simulated a log-normal random variable having mean $\begin{pmatrix} e^{\mu_1+\sigma^2/2} \\ e^{\mu_2+\sigma^2/2} \end{pmatrix}$ and median $\begin{pmatrix} e^{\mu_1} \\ e^{\mu_2} \end{pmatrix}$. The X had the same normal settings as in the previous text.

Similar to the normal cases, we tabulated the bias, mean square error, confidence interval width, empirical size and power in Table II. Table II illustrates that the hybrid method resulted in a slightly lower type I error rate than the meta-analysis t-test but the empirical power of hybrid method was higher as compared with the classic meta-analysis t-test. Under the null hypothesis, the bias of the estimate based

Table I. Simulated bias, mean square error, 95% confidence interval width, empirical size and power at a significance level 0.05 with 1000 replicates under normal distribution.

Under null hypothesis

			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical size	Bias*	MSE	CI width	Empirical size	Bias*	MSE	CI width
240	35	0.1	0.0500	0.0009	0.0068	0.3277	0.0580	0.0002	0.0065	0.3182
		0.2	0.0440	0.0040	0.0060	0.3109	0.0460	0.0031	0.0057	0.3021
		0.3	0.0540	−0.0009	0.0056	0.2932	0.0560	−0.0002	0.0054	0.2848
		0.4	0.0430	−0.0017	0.0048	0.2732	0.0490	−0.0007	0.0045	0.2650
		0.5	0.0530	0.0024	0.0042	0.2505	0.0510	0.0025	0.0040	0.2436
		0.6	0.0370	−0.0010	0.0030	0.2258	0.0410	−0.0010	0.0029	0.2192
		0.7	0.0450	−0.0016	0.0025	0.1973	0.0450	−0.0018	0.0024	0.1917
		0.8	0.0410	−0.0013	0.0016	0.1622	0.0430	−0.0017	0.0015	0.1576
		0.9	0.0430	−0.0005	0.0008	0.1153	0.0450	−0.0006	0.0008	0.1121

Under alternative hypothesis

			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical power	Bias*	MSE	CI width	Empirical power	Bias*	MSE	CI width
240	35	0.1	0.2390	0.0002	0.0076	0.3288	0.2470	−0.0010	0.0072	0.3185
		0.2	0.2490	0.0003	0.0066	0.3113	0.2640	−0.0010	0.0063	0.3019
		0.3	0.2570	0.0001	0.0056	0.2933	0.2750	0.0009	0.0053	0.2846
		0.4	0.3180	0.0013	0.0048	0.2734	0.3200	0.0008	0.0046	0.2656
		0.5	0.3690	0.0035	0.0041	0.2506	0.4050	0.0040	0.0038	0.2434
		0.6	0.4170	0.0002	0.0034	0.2260	0.4290	−0.0005	0.0033	0.2197
		0.7	0.5060	0.0010	0.0027	0.1969	0.5260	0.0014	0.0025	0.1916
		0.8	0.6780	0.0008	0.0017	0.1622	0.7130	0.0006	0.0016	0.1577
		0.9	0.9270	−0.0014	0.0009	0.1153	0.9340	−0.0015	0.0008	0.1121

MSE, mean square error; CI, confidence interval.

Table II. Simulated bias, mean square error, 95% confidence interval width, empirical size and power at a significance level 0.05 with 1000 replicates under log-normal distribution.

Under null hypothesis

			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical size	Bias*	MSE	CI width	Empirical size	Bias*	MSE	CI width
240	35	0.1	0.0270	−0.0015	0.0072	0.3583	0.0380	0.0008	0.0300	0.7008
		0.2	0.0350	0.0040	0.0068	0.3397	0.0440	0.0030	0.0273	0.6685
		0.3	0.0310	0.0031	0.0060	0.3183	0.0500	0.0024	0.0258	0.6441
		0.4	0.0400	0.0013	0.0054	0.2962	0.0570	−0.0010	0.0249	0.6067
		0.5	0.0450	0.0010	0.0044	0.2704	0.0420	−0.0034	0.0226	0.5717
		0.6	0.0400	−0.0017	0.0037	0.2439	0.0540	−0.0038	0.0192	0.5281
		0.7	0.0500	0.0022	0.0029	0.2110	0.0520	0.0048	0.0147	0.4714
		0.8	0.0450	0.0021	0.0019	0.1735	0.0530	0.0040	0.0107	0.3954
		0.9	0.0400	0.0005	0.0009	0.1231	0.0440	0.0039	0.0054	0.2870

Under alternative hypothesis (median)

			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical power	Bias*	MSE	CI width	Empirical power	Bias*	MSE	CI width
240	35	0.1	0.1440	−0.0216	0.0637	1.0298	0.1270	0.1302	0.2485	1.9223
		0.2	0.1890	−0.0051	0.0585	0.9757	0.1460	0.1659	0.2545	1.8647
		0.3	0.2330	0.0012	0.0516	0.9163	0.1810	0.1839	0.2504	1.7879
		0.4	0.2260	−0.0094	0.0443	0.8544	0.1800	0.1658	0.2064	1.6996
		0.5	0.2630	−0.0172	0.0356	0.7801	0.1940	0.1532	0.1863	1.5898
		0.6	0.3170	−0.0126	0.0301	0.7048	0.2530	0.1764	0.1916	1.4650
		0.7	0.4250	−0.0118	0.0236	0.6185	0.3120	0.1879	0.1498	1.3164
		0.8	0.5420	−0.0188	0.0152	0.5088	0.3790	0.1547	0.1061	1.1242
		0.9	0.8240	−0.0158	0.0096	0.3710	0.6270	0.1780	0.0787	0.8234

MSE, mean square error; CI, confidence interval.

Statistics in Medicine

on the hybrid statistic was similar to that based on meta-analysis t-test. However, the estimate derived from hybrid test statistic had significant small mean square error and narrow confidence interval under the null and the alternative. For results in Table II, we used median as the location parameter. We also simulated results based on mean as the location parameter. From this setting, we had the same observations as using median except for bias under the alternative. Using mean as the location parameter, the bias simulated from the estimate based on hybrid test statistic was generally larger than that based on meta-analysis t-test statistic. The observation was due to the fact that our estimate based on hybrid statistic used median as the location parameter whereas meta-analysis t-test statistic used mean as the location parameter. If we compare the mean square error and confidence interval width derived from estimates based on hybrid test statistic using median as the location parameter to those from estimates based on meta-analysis t-test statistic using mean as the location parameter, the proposed method resulted in smaller mean square error and narrower confidence interval (The results relating to simulations used mean as the location parameter for log-normal were not shown in tables).

4.2. Discrete case

4.2.1. Multinomial distribution. For simulating possible discrete observations in the study such as ETROP, we selected multinomial distribution. On the basis of ETROP study, we set the similar conditions on the number of categories and sample in each category as in ETROP study. First, we generated random realizations from bivariate and univariate normal distributions for matched and unmatched studies as in previous cases, respectively. Second, we defined cutoffs on the basis of the data from ETROP study and under null hypothesis, and we assigned the same cutoffs on treatment and control in matched and unmatched study. Then, we computed which category normal realizations fall into and assigned the cutoff value to each normal realization. In the simulation, we assigned 24 and 12 categories on matched and unmatched components, respectively. Under the alternative hypothesis, we assigned 24 on matched, 11 and 13 on unmatched treatment and control components, respectively. We tabulate the results in Table III. Table III indicated that the hybrid method has lower type I error rate than the classic meta-analysis t-test. The empirical powers of the classic meta-analysis t-test were higher than those of the hybrid test when the correlations were greater or equal to 0.3. Under the null hypothesis, the simulated bias and mean square error were similar, but the confidence interval derived from the estimate based on hybrid statistic was narrower than that from meta-analysis t-test statistic. Under the alternative, the underlying location parameter was unknown, and we were unable to simulate the bias and mean square error.

4.3. Outlier case

In practice, we may observe outliers. In ETROP study, an extremely large number was assigned if the eye is blind compared with a value of visual acuity with normal vision.

From multinomial case, we assigned a larger number in some categories to represent outlier cases. In this simulation, we set 10% outliers from normal case by assigning an artificial larger number (500) to replace the originally simulated value. Whereas parametric method cannot detect any case given significant level 0.05, hybrid method can detect. This means that hybrid method is robust and more appropriate to apply on data with outlier cases. For the estimate based on hybrid statistic, we still had meaningful estimate, and the bias, mean square error and confidence interval width were under control. However, the estimates based on meta-analysis t-test statistic become useless. We tabulate the results in Table IV.

5. Application of the hybrid test statistic

The 6-year ETDRS acuity outcomes were measured for patients in the trial [10]. We applied our method to both type I and type II eyes to assess the treatment effects. We applied our test statistic to the ETDRS acuity score in logarithm of the minimum angle of resolution (logMAR) unit to both types of eyes. LogMAR measures visual acuity loss, which is a log10 transformation of the acuity fraction, $\text{logMAR} = -\log10(20/x)$. Increased value of logMAR indicates vision loss, and decreased value of logMAR denotes normal or better visual acuity. We performed a battery of statistical tests and plots such as Kolmogorov–Smirnov test (p-values were less than 0.0001), histograms (Figures 1 and 2) and Box–Cox plots (not shown). The results clearly indicated that visual acuity scores in logMAR scale were inconsistent with a normally distributed random variable.

Table III. Simulated bias, mean square error, 95% confidence interval width, empirical size and power at a significance level 0.05 with 1000 replicates under multinomial distribution.

			Under null hypothesis							
			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical size	Bias*	MSE	CI width	Empirical size	Bias*	MSE	CI width
240	35	0.1	0.0420	0.0013	0.0096	0.3927	0.0510	−0.0034	0.0235	0.6105
		0.2	0.0480	0.0020	0.0085	0.3719	0.0600	0.0057	0.0220	0.5858
		0.3	0.0340	0.0038	0.0071	0.3479	0.0550	0.0042	0.0195	0.5582
		0.4	0.0440	−0.0054	0.0062	0.3197	0.0450	−0.0095	0.0177	0.5289
		0.5	0.0510	−0.0022	0.0053	0.2857	0.0450	0.0023	0.0153	0.4936
		0.6	0.0490	−0.0018	0.0044	0.2546	0.0580	−0.0044	0.0148	0.4546
		0.7	0.0420	0.0010	0.0028	0.2139	0.0470	0.0064	0.0108	0.4087
		0.8	0.0740	−0.0008	0.0016	0.1681	0.0470	0.0002	0.0077	0.3495
		0.9	0.0300	−0.0013	0.0004	0.1080	0.0420	0.0000	0.0047	0.2672
			Under alternative hypothesis							
			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical power	Bias*	MSE	CI width	Empirical power	Bias*	MSE	CI width
240	35	0.1	0.4720	.	.	0.3852	0.4310	.	.	0.5954
		0.2	0.4610	.	.	0.3672	0.4410	.	.	0.5720
		0.3	0.4580	.	.	0.3432	0.5050	.	.	0.5456
		0.4	0.4530	.	.	0.3212	0.5220	.	.	0.5176
		0.5	0.4650	.	.	0.2918	0.5920	.	.	0.4860
		0.6	0.5140	.	.	0.2649	0.6250	.	.	0.4509
		0.7	0.5760	.	.	0.2286	0.7330	.	.	0.4065
		0.8	0.5600	.	.	0.1883	0.8320	.	.	0.3524
		0.9	0.6170	.	.	0.1439	0.9710	.	.	0.2806

MSE, mean square error; CI, confidence interval.

Table IV. Simulated bias, mean square error, 95% confidence interval width, empirical size and power at a significance level 0.05 with 1000 replicates under normal distribution with 10% outliers.

			Under null hypothesis							
			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical size	Bias*	MSE	CI width	Empirical size	Bias*	MSE	CI width
240	35	0.1	0.0120	0.0014	0.0083	0.4515	0	−1.9785	3079.4223	1011.1019
		0.2	0.0160	−0.0033	0.0078	0.4302	0	3.5624	3267.4405	1011.6055
		0.3	0.0220	−0.0009	0.0071	0.4049	0	−0.8549	3348.5636	1010.5189
		0.4	0.0110	0.0003	0.0057	0.3812	0	0.7847	3108.1675	1010.9635
		0.5	0.0100	0.0013	0.0047	0.3509	0	0.9139	3058.6077	1011.1564
		0.6	0.0140	−0.0023	0.0041	0.3177	0	−1.9407	3149.2831	1011.4315
		0.7	0.0130	0.0038	0.0031	0.2793	0	0.1731	3100.8582	1012.9987
		0.8	0.0120	0.0043	0.0022	0.2316	0	1.2377	2945.8480	1011.6689
		0.9	0.0180	−0.0010	0.0011	0.1652	0	−3.2054	3178.5051	1009.5425
			Under alternative hypothesis							
			T_{Hybrid}				T^*_{Classic}			
n_1	n_2, n_3	ρ	Empirical power	Bias*	MSE	CI width	Empirical power	Bias*	MSE	CI width
240	35	0.1	0.0780	−0.0042	0.0086	0.4524	0	3.0599	2662.5693	1011.6090
		0.2	0.0890	−0.0015	0.0075	0.4303	0	−1.6420	2867.5463	1012.1438
		0.3	0.0980	0.0024	0.0060	0.4070	0	−0.4638	2935.1688	1010.4602
		0.4	0.1150	−0.0005	0.0055	0.3811	0	−1.9324	2979.9895	1011.8693
		0.5	0.1290	−0.0059	0.0048	0.3511	0	1.1555	2793.1063	1011.8664
		0.6	0.1600	−0.0036	0.0038	0.3184	0	1.5052	2949.5547	1010.8637
		0.7	0.2310	−0.0008	0.0030	0.2798	0	0.7260	3134.0223	1012.6820
		0.8	0.3340	−0.0041	0.0021	0.2314	0	0.0096	2793.7188	1013.0948
		0.9	0.6640	−0.0031	0.0011	0.1657	0	−1.7452	3116.9506	1011.7281

MSE, mean square error; CI, confidence interval.

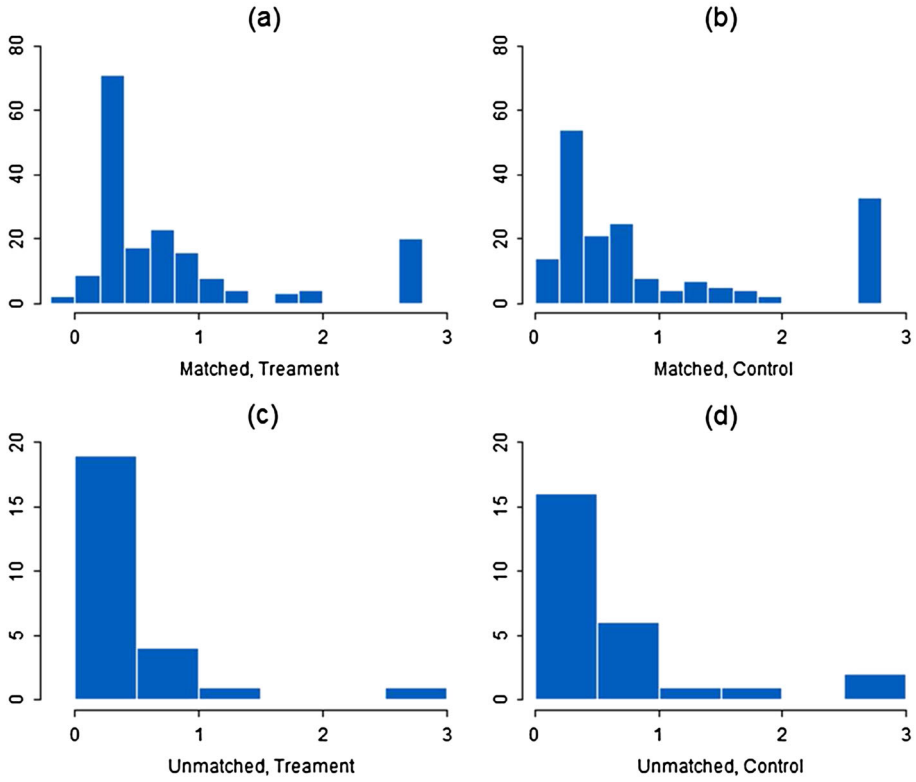

Figure 1. Histograms of the logarithm of the minimum angle of resolution from type I eyes. (a) Matched eyes with treatment, (b) matched eyes with control, (c) unmatched eye with treatment and (d) unmatched eyes with control.

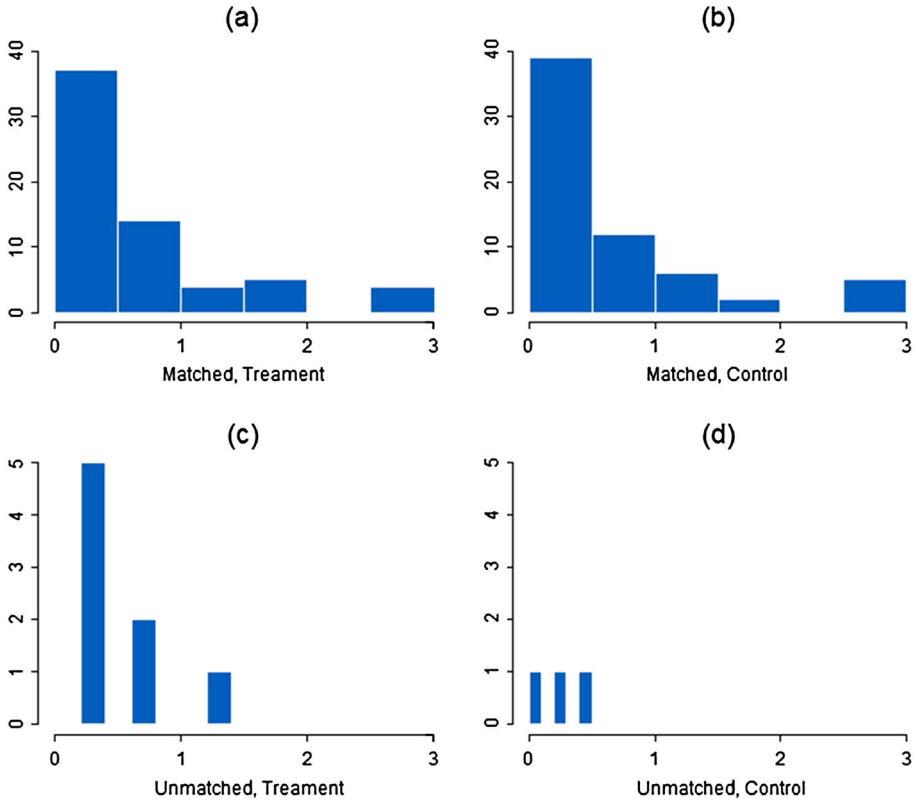

Figure 2. Histograms of the logarithm of the minimum angle of resolution from type II eyes. (a) Matched eyes with treatment, (b) matched eyes with control, (c) unmatched eye with treatment and (d) unmatched eyes with control.

For the matched type I eyes, the treatment effect was −0.097. On the basis of the estimate of the variance (0.001) and the normality assumption of the hybrid test statistic, the 95% confidence interval for the treatment effect was (−0.171, −0.023). This 95% confidence interval was slightly different from the 95% confidence interval (−0.151, −0.003) from the Hodges–Lehmann estimate based on the Wilcoxon signed rank test. For the unmatched type I eyes, the treatment effect was −0.194, the estimated variance was 0.006, the 95% confidence interval derived from the estimate based on hybrid test statistic was (−0.267, −0.120), and the 95% confidence interval was (−0.301, 0.000) by the Hodges–Lehmann estimate based on Wilcoxon rank sum test. By using our estimate based on hybrid test statistic, the combined treatment effect for type I eyes was −0.1156, and the 95% confidence interval was (−0.182, −0.120). No direct Wilcoxon type interval was available for the combined case. We may treat our hybrid method as a Hodges–Lehmann type estimate based on Wilcoxon statistics for combining treatment effect from matched and unmatched studies. The z-value from the estimate based on the hybrid test statistic was −3.429 (p-value = 0.0006).

We have performed similar calculations for type II eyes. The treatment effect for the matched eyes was 0.052 with variance 0.002, and the 95% confidence interval was (−0.025, 0.129) by estimate derived from the hybrid test statistic, whereas the 95% confidence interval was (0, 0.154) for estimate derived from the Wilcoxon signed rank test. For the unmatched eyes, the treatment effect was 0.011 with variance 0.144. The 95% confidence interval was (0.030, 0.188) by estimate derived from the hybrid test statistic. The 95% confidence interval was (−0.294, 1.194) for estimate derived from the Wilcoxon rank sum test, which was much wider because of small number of observations in the unmatched case. Using the estimate based on hybrid test statistic, we obtained the combined treatment effect of 0.053, and the 95% confidence interval was (−0.024, 0.129). Again, there was no existing confidence interval derived from estimate based on Wilcoxon type statistic for the combined case. The z-value derived from the estimate based on hybrid test statistic was 1.343 (p-value = 0.179).

6. Concluding remarks

The classic nonparametric procedures corresponding to the paired t-test and the two sample t-test are the Wilcoxon signed rank test and the Wilcoxon rank sum test, respectively. We notice that the Wilcoxon signed rank test for matched pairs and the Wilcoxon rank sum test for the unmatched pairs were in different scales of measurements. We were unable to directly combine the Wilcoxon signed rank test and the Wilcoxon rank sum test as in (4), and the traditional meta-analysis average based on direct weighting of both test statistics is not meaningful. However, the treatment effects measured by the medians based on the Hodges–Lehmann estimates from the test statistics are in the same scale. So, we can use the estimate based on our hybrid test statistic in combing the treatment effects from matched and unmatched studies directly. Our simulations showed that the estimate based on the hybrid statistic had smaller bias and mean square error as well as narrower confidence interval. The Hodges–Lehmann estimates are asymptotically normally distributed [13]. Furthermore, the variance estimates are consistent estimates of the underlying variances of the Hodges–Lehmann estimates [18].Therefore, the hybrid test statistic proposed in this article would be asymptotically normally distributed after proper scaling.

Another advantage of using the proposed hybrid statistic for combining treatment effects from both matched and unmatched studies is that the hybrid method can be applied when there are outliers in the measurements. For example, in our application presented in the last section, the results for the classic meta-analysis t-test were not available because of outlier values from the study. Our simulation results in Table IV also indicated that the failure of the meta-analysis t-test when outliers were presented. However, the hybrid test preserved reasonable power and controlled the test size. In our simulation study, as in Zhou *et al.* [19], we used the nominal sizes and their corresponding asymptotic critical values in comparing the hybrid and meta-analysis t-test statistics. The empirical sizes of hybrid statistics were very close to that of the meta-analysis t-test under normal setting. Under other settings, we observed that the hybrid test statistic had empirical sizes less or equal to that of the classical meta-analysis t-test based on the asymptotic critical value. Therefore, the actual powers of the hybrid test would be higher than the empirical powers in Tables II, III and IV because of our conservative approach in comparing the powers.

The proposed hybrid statistic does not take into account covariates directly. In randomized clinical trials, primary analysis for main covariates is usually performed through stratifications and/or regression analysis. The proposed hybrid method is still useful for stratification analysis [10]. Although our hybrid statistic is robust to outliers, more care is needed for any other estimates and test statistics in handling missing values in clinical trials. In our application, there were very few missing values.

Acknowledgements

The study was supported by Cooperative Agreements (5U10 EY12471 and 5U10 EY12472) with the National Eye Institute of the National Institutes of Health, U.S. Department of Health and Human Services, Bethesda, Maryland. We also thank Charles Lai for his editing.

References

1. Hardy RJ, Davis BR, Tung B, Palmer E. Statistical considerations in terminating the multicenter trial of cryotherapy of retinopathy of prematurity. *Controlled Clinical Trials* 1991; **12**:1408–1416.
2. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: Snellen acuity and structural outcome at 5 $^1/_2$ years. *Archives of Ophthalmology* 1996; **114**:417–424.
3. Good WV, Hardy RJ. The multicenter study of Early Treatment for Retinopathy of Prematurity (ETROP). *Ophthalmology* 2001; **108**:1013–1014.
4. Early Treatment for Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: Results of the early treatment for retinopathy of prematurity randomized trial. *Archives of Ophthalmology* 2003; **121**:1684–1694.
5. Early Treatment for Retinopathy of Prematurity Cooperative Group. Multicenter trial of early treatment for retinopathy of prematurity: study design. *Controlled Clinical Trials* 2004; **25**:311–325.
6. Shutton AJ, Abrams KRJDR. *Methods for Meta-Analysis in Medical Research*. Wiley: 2000.
7. Seber F, Lee AJ. *Linear Regression Analysis*, 2nd ed. Wiley: New York, 2003.
8. Le Cessie S, Nagelkerke N, Rosendaal FR. Combining matched and unmatched control groups in case-control studies. *American Journal of Epidemiology* 2008; **168**:1204–1210.
9. Hardy RJ, Palmer EA, Dobson V, *et al*. Risk analysis of prethreshold retinopathy of prematurity. *Archives of Ophthalmology* 2003; **121**:1697–1701.
10. The Early Treatment for Retinopathy of Prematurity Cooperative Group. Final visual acuity results in the early treatment for retinopathy of prematurity study. *Archives of Ophthalmology* 2010; **128**:663–671.
11. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. Wiley: New York, 1999.
12. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945; **1**:80–83.
13. Hodges JL, Lehmann EL. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 1963; **34**:598–611.
14. Walsh JE. Some significance tests for the median which are valid under very general conditions. *The Annals of Mathematical Statistics* 1949; **20**:64–81.
15. Rosenthal R. Combining results of independent studies. *Psychological Bulletin* 1978; **85**:185–193.
16. McKean JW, Schrader R. A comparison of methods for studentizing the sample median. *Communications in Statistics, Ser. B* 1984; **13**:751–773.
17. Price RM, Bonett DG. Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation* 2001; **68**:295–305.
18. Lehmann EL. Nonparametric confidence intervals for a shift parameter. *Annals of Mathematical Statistics* 1963; **34**:1507–1512.
19. Zhou XH, Gao SJ, Hui SL. Methods for comparing the means of two independent log-normal samples. *Biometrics* 1997; **53**:1129–1135.