Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Chiba, J Biom Biostat 2015, 6:3
DOI: 10.4172/2155-6180.1000244

Research Article Open Access

Exact Tests for the Weak Causal Null Hypothesis on a Binary Outcome in Randomized Trials

Yasutaka Chiba*

Clinical Research Center, Kinki University Hospital, Japan

Abstract

There are two principal exact tests for evaluation of data in two-by-two contingency tables: the tests of Fisher and Barnard. The latter cannot be a hypothesis test for the causal null hypothesis unless exchangeability can be assumed. Fisher's exact test is a hypothesis test for the sharp causal null hypothesis (i.e., that there is no effect for all individuals), but not for the weak causal null hypothesis (i.e., that the true risk difference is zero). Rejection of the sharp causal null hypothesis does not mean that the weak causal null hypothesis is rejected (i.e., that the true risk difference is not zero). In this article, we provide exact tests for the weak causal null hypothesis, in the absence of any assumption, in the context of randomized trials. Using the concept of principal stratification, which considers four types of subjects to define four principal strata, we derive an unconditional exact test, for which neither marginal total is fixed, and a conditional exact test, for which one marginal total is fixed. In addition, we show that Fisher's exact test can be a hypothesis test for the weak causal null hypothesis when monotonicity can be assumed. The derived exact tests are extended to hypothesis testing for non-inferiority trials and to construct confidence intervals linking to the exact tests. The derived exact tests and confidence intervals are illustrated using data from two clinical trials.

Keywords: Complete (simple) randomization; Conditional and unconditional exact tests; Exchangeability; Monotonicity; Potential outcome; Principal stratification

Introduction

We consider a randomized trial, in which subjects are assigned randomly to receive one of two experimental treatments, and each subject is classified as either a responder or a non-responder. In such a setting, data are summarized in a two-by-two contingency table, and hypothesis testing is performed to test the equality of the response proportions of the two groups.

Ninety years ago, Fisher [1,2] developed an exact test, which is often referred to as Fisher's exact test. This test is a hypothesis test for the sharp, but not the weak, causal null hypothesis. Because rejection of the sharp causal null hypothesis does not imply that the weak causal null hypothesis is rejected, the true risk difference may still be zero even when the sharp causal null hypothesis is rejected by Fisher's exact test.

Twenty years later, Barnard [3-5] developed another exact test, which is sometimes referred to as Barnard's exact test. This test utilizes the product of two binomial probabilities, and it has advantages over Fisher's exact test in that it can be more powerful for moderate to small samples [6,7]. Nevertheless, Barnard's exact test cannot be a hypothesis test for the causal null hypothesis unless exchangeability can be assumed.

In this article, we provide exact tests for the weak causal null hypothesis without requiring any assumption (such as exchangeability) in the context of randomized trials, using the concept of principal stratification. To our knowledge, such an exact test has not been developed. First, an unconditional exact test for which neither marginal total is fixed is derived; then, a conditional exact test for which one marginal total is fixed is derived. In addition, we show that Fisher's exact test can be a hypothesis test for the weak causal null hypothesis when monotonicity can be assumed. The derived exact tests are extended to hypothesis testing for non-inferiority trials and to construct confidence intervals linking to the exact tests. The derived exact tests and confidence intervals are illustrated using data from two clinical trials.

Notation and Principal Stratification Approach

Throughout this article, we denote X as the assigned treatment; X=1 if a subject was assigned to the treatment group, and X=0 if assigned to the control group. Y denotes the binary outcome; Y=1 if the event occurred, and Y=0 if it did not. The results from a randomized trial are summarized in a two-by-two contingency table as shown in Table 1, where a, b, c, d, and n are the numbers of subjects.

For each subject, it is also possible to consider the potential outcomes [8-10], which correspond to the outcomes of the subject had he/she been in the other group of the trial. $Y(x)$ denotes the potential outcome for each subject under $X=x$. Then, $\Pr(Y(1)=1)$ represents a potential response proportion if all subjects are assigned to the treatment group, and $\Pr(Y(0)=1)$ represents a potential response proportion if all subjects are assigned to the control group. Using the potential outcome, the null hypothesis that the potential response proportions of each group are equal can be defined as

$$H_0: \Pr(Y(1)=1)=\Pr(Y(0)=1).$$

	Event		
Group	**Occurred (*Y*=1)**	**Not occurred (*Y*=0)**	**Total**
Treatment (*X*=1)	*a*	*b*	*a+b*
Control (*X*=0)	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*n*

Table 1: Two-by-two contingency table obtained from a randomized trial, where *a*, *b*, *c*, *d*, and *n* indicate the numbers of subjects.

***Corresponding author:** Yasutaka Chiba, Clinical Research Center, Kinki University Hospital, Japan, Tel: +81-72-366-0221, Fax: +81-72-368-1193; E-mail: chibay@med.kindai.ac.jp

Received July 27, 2015; **Accepted** August 18, 2015; **Published** August 25, 2015

Citation: Chiba Y (2015) Exact Tests for the Weak Causal Null Hypothesis on a Binary Outcome in Randomized Trials. J Biom Biostat 6: 244. doi:10.4172/2155-6180.1000244

This null hypothesis is referred to as the weak causal null hypothesis [11], and implies that two different treatment statuses from one population are compared.

Here, we apply the principal stratification approach [12]. This approach considers the following four types of subjects to define the four principal strata:

(i) Individuals who would suffer the event regardless of the assigned treatment group; i.e., $(Y(1), Y(0))=(1, 1)$.

(ii) Individuals who would suffer the event if assigned to the treatment group, but would not suffer the event if assigned to the control group; i.e., $(Y(1), Y(0))=(1, 0)$.

(iii) Individuals who would not suffer the event if assigned to the treatment group, but would suffer the event if assigned to the control group; i.e., $(Y(1), Y(0))=(0, 1)$.

(iv) Individuals who would not suffer the event regardless of the assigned treatment group; i.e., $(Y(1), Y(0))=(0, 0)$.

These four types are summarized in Table 2, and all subjects belong to one of these four types.

Let n_{st} denote the number of subjects with $(Y(1), Y(0))=(s, t)$, where $s, t=0, 1$. Although the value of n_{st} cannot be determined from the observed data, we can nevertheless express the weak causal null hypothesis by using n_{st}. If all subjects are assigned to the treatment group ($X=1$), $\Pr(Y(1)=1)=(n_{11}+n_{10})/n$ because only subjects with type (i) or (ii) would suffer the event (Table 2). Likewise, if all subjects are assigned to the control group ($X=0$), $\Pr(Y(0)=1)=(n_{11}+n_{01})/n$ because only subjects with type (i) or (iii) would suffer the event. Thus, using the concept of the principal stratification, the weak causal null hypothesis can be expressed as $n_{10}=n_{01}$ from $\Pr(Y(1)=1)=\Pr(Y(0)=1)$.

Proposed Exact Tests

Using the above notation, we derive an unconditional exact test for the weak causal null hypothesis $n_{10}=n_{01}$ in the context of randomized trials with complete (or equally simple) randomization, and apply it to derive a conditional exact test.

Unconditional exact test

When the random assignment is conducted by the ratio of 1:r, we assume that subjects are assigned as in Table 3 under the null hypothesis; i.e., of the n_{st} subjects, $n_{st,1}$ subjects are assigned to the treatment group ($X=1$) with the probability of $1/(1+r)$ and $n_{st,0}$ subjects are assigned to the control group ($X=0$) with the probability of $r/(1+r)$. As each subject is independently assigned, we form the following probability:

$$\begin{aligned}&\Pr(N_{11,1}=n_{11,1}, N_{10,1}=n_{10,1}, N_{01,1}=n_{01,1}, N_{00,1}=n_{00,1})\\&=\prod_{s=0}^{1}\prod_{t=0}^{1}\Pr(N_{st,1}=n_{st,1})\\&=\prod_{s=0}^{1}\prod_{t=0}^{1}\binom{n_{st}}{n_{st,1}}\left(\frac{1}{1+r}\right)^{n_{st,1}}\left(\frac{r}{1+r}\right)^{n_{st,0}},\end{aligned}\quad (1)$$

where $\binom{j}{k}={}_jC_k=\frac{j!}{k!(j-k)!}$, and the following set of conditions is required:

Set of conditions 1:

$$n_{10}=n_{01}, \sum_{s=0}^{1}\sum_{t=0}^{1}n_{st}=n, \begin{cases}n_{11}\le a+c\\ n_{10}\le a+d\\ n_{01}\le b+c\\ n_{00}\le b+d\end{cases} and \begin{cases}n_{11}+n_{10}\le a+c+d=n-b\\ n_{11}+n_{01}\le a+b+c=n-d\\ n_{00}+n_{10}\le a+b+d=n-c\\ n_{00}+n_{01}\le b+c+d=n-a\end{cases}$$

The first condition is the null hypothesis and the second is the total number of subjects. The last two conditions are needed such that the numbers of subjects in principal strata under the null hypothesis, $(n_{11}, n_{10}, n_{01}, n_{00})$, do not contradict the observed data in Table 1; i.e., Table 3 is equal to Table 1 under at least one combination of $n_{st,1}$ and $n_{st,0}$. If $(n_{11}, n_{10}, n_{01}, n_{00})$ does contradict the observed data, subjects in the principal strata can no longer be the same sample as the subjects in the observed data. These conditions are derived from Tables 1 and 3 as follows; *e.g.*, $n_{11}\le a+c$ is derived from

$$\begin{aligned}&a+c\\&=(n_{11,1}+n_{10,1})+(n_{11,0}+n_{01,0})\\&=n_{11}+n_{10,1}+n_{01,0},\end{aligned}$$

and $n_{11}+n_{10}\le a+c+d$ is derived from

$$\begin{aligned}&a+c+d\\&=(n_{11,1}+n_{10,1})+(n_{11,0}+n_{01,0})+(n_{00,0}+n_{10,0})\\&=n_{11}+n_{10}+n_{01,0}+n_{00,1}.\end{aligned}$$

The other inequalities are derived in a similar manner.

Here, we focus on the risk difference as the effect measure. The risk difference estimated from the observed data is

$$\mathrm{RD}_{\mathrm{O}}:=\frac{a}{a+b}-\frac{c}{c+d}$$

from Table 1, and the risk difference under the null hypothesis is

$$\mathrm{RD}_{\mathrm{N}}:=\frac{n_{11,1}+n_{10,1}}{n_{11,1}+n_{10,1}+n_{01,1}+n_{00,1}}-\frac{n_{11,0}+n_{01,0}}{n_{11,0}+n_{10,0}+n_{01,0}+n_{00,0}}$$

from Table 3. We consider only the case of $\mathrm{RD}_{\mathrm{O}}\le 0$ in this article, but the following methods can easily be applied to the case of $\mathrm{RD}_{\mathrm{O}}\ge 0$. For $\mathrm{RD}_{\mathrm{O}}\le 0$, the one-sided p-value is defined as the probability that RD_{N} is equal to or smaller than RD_{O} if the same trial is conducted repeatedly under the null hypothesis. Therefore, Equation (1) yields the following one-sided p-value under a combination of $(n_{11}, n_{10}, n_{01}, n_{00})$ satisfying

Type	Principal stratum	Event under the treatment group	Event under the control group
(i)	$(Y(1), Y(0))=(1, 1)$	Yes	Yes
(ii)	$(Y(1), Y(0))=(1, 0)$	Yes	No
(iii)	$(Y(1), Y(0))=(0, 1)$	No	Yes
(iv)	$(Y(1), Y(0))=(0, 0)$	No	No

Table 2: Principal strata: "Yes" denotes that a subject would suffer the event, and "No" denotes that a subject would not suffer the event.

	Event		
Group	Occurred ($Y=1$)	Not occurred ($Y=0$)	Total
Treatment ($X=1$)	$n_{11,1}+n_{10,1}$	$n_{00,1}+n_{01,1}$	$n_{11,1}+n_{10,1}+n_{01,1}+n_{00,1}$
Control ($X=0$)	$n_{11,0}+n_{01,0}$	$n_{00,0}+n_{10,0}$	$n_{11,0}+n_{10,0}+n_{01,0}+n_{00,0}$
Total	$n_{11}+n_{10,1}+n_{01,0}$	$n_{00}+n_{01,1}+n_{10,0}$	n

Table 3: Two-by-two contingency table with the numbers for the four types of subjects defining the four principal strata.

set of conditions 1:

$$p_{n_{11},n_{10},n_{01},n_{00}} = \sum_{n_{11,1}=0}^{n_{11}} \sum_{n_{10,1}=0}^{n_{10}} \sum_{n_{01,1}=0}^{n_{01}} \sum_{n_{00,1}=0}^{n_{00}} I(z) \prod_{s=0}^{1} \prod_{t=0}^{1} \binom{n_{st}}{n_{st,1}} \left(\frac{1}{1+r}\right)^{n_{st,1}} \left(\frac{r}{1+r}\right)^{n_{st,0}}, \quad (2)$$

where $I(z)=1$ if $z \le 0$ and $I(z)=0$ if $z>0$ with $z := \mathrm{RD}_N\text{-}\mathrm{RD}_O$. Note that for cases in which one denominator in RD_N is 0, we set the indicator to $I(z)=1$. Although this setting of the indicator yields larger p-values than setting to $I(z)=0$, the substantial effect will be trivial because the probability that either $n_{st,1}$ or $n_{st,0}$ is 0 for all s and t is very small.

Unfortunately, this calculation of the p-value yields plural p-values corresponding to the number of combinations of $(n_{11}, n_{10}, n_{01}, n_{00})$, and then we cannot yield a p-value immediately. A method to deal with such a problem is to calculate the p-values for all possible combinations of $(n_{11}, n_{10}, n_{01}, n_{00})$ and choose the maximum value [3]. Such a method may make the result of the hypothesis testing conservative. Using this method, we define the unconditional exact p-value based on the principal stratification as follows:

$$p = sup\{p_{n_{11},n_{10},n_{01},n_{00}} : (n_{11}, n_{10}, n_{01}, n_{00}) \text{ satisfying set of conditions 1}\}.$$

We note that neither marginal total is fixed for the unconditional exact test.

Conditional exact test

While the unconditional exact test does not fix the numbers of subjects assigned to the two groups, the conditional exact test does. Therefore, we consider a conditional probability on $\Sigma_s\Sigma_t N_{st,1}=\Sigma_s\Sigma_t n_{st,1}$ $(=a+b)$ instead of Equation (1). This conditional probability can be expressed as

$$\begin{aligned}
&\Pr\left(N_{11,1}=n_{11,1}, N_{10,1}=n_{10,1}, N_{01,1}=n_{01,1}, N_{00,1}=n_{00,1} \,\middle|\, \sum_{s=0}^{1}\sum_{t=0}^{1} N_{st,1} = \sum_{s=0}^{1}\sum_{t=0}^{1} n_{st,1}\right) \\
&= \prod_{s=0}^{1}\prod_{t=0}^{1} \Pr(N_{st,1}=n_{st,1}) \Big/ \Pr\left(\sum_{s=0}^{1}\sum_{t=0}^{1} N_{st,1} = a+b\right) \\
&= \prod_{s=0}^{1}\prod_{t=0}^{1} \binom{n_{st}}{n_{st,1}} \Big/ \binom{n}{a+b},
\end{aligned}$$

where the following conditions are required:

Set of conditions 2:

Set of conditions 1 plus $\sum_{s=0}^{1}\sum_{t=0}^{1} n_{st,1} = a+b$.

Consequently, we can define the conditional exact p-value based on the principal stratification as follows:

$$p = sup\{p_{n_{11},n_{10},n_{01},n_{00}} : (n_{11}, n_{10}, n_{01}, n_{00}) \text{ satisfying set of conditions 2}\}$$

with

$$p_{n_{11},n_{10},n_{01},n_{00}} = \sum_{n_{11,1}=0}^{n_{11}} \sum_{n_{10,1}=0}^{n_{10}} \sum_{n_{01,1}=0}^{n_{01}} \sum_{n_{00,1}=0}^{n_{00}} I(z) \prod_{s=0}^{1} \prod_{t=0}^{1} \binom{n_{st}}{n_{st,1}} \Big/ \binom{n}{a+b}. \quad (3)$$

Other Exact Tests

Here, we discuss assumptions for Fisher's and Barnard's exact tests being hypothesis tests for the weak causal null hypothesis.

Fisher's exact test

First, we show that Fisher's exact test is a special case of the conditional exact test given here, with the null hypothesis of $n_{10}=n_{01}=0$. In this case, set of conditions 1 is $n_{10}=n_{01}=0$, $n_{11}+n_{00}=n$, $n_{11} \le a+c$ and $n_{00} \le b+d$, and thus $(n_{11}, n_{10}, n_{01}, n_{00})=(a+c, 0, 0, b+d)$. In addition, under $\Sigma_s\Sigma_t n_{st,1}=n_{11,1}+n_{00,1}=a+b$ in set of conditions 2, $n_{11,1} \ge a\text{-}d$ because $b+d=n_{00,1}+n_{00,0} \ge n_{00,1}=a+b\text{-}n_{11,1}$, and

$$\begin{aligned}
z &= \mathrm{RD}_N - \mathrm{RD}_O \\
&= \left(\frac{n_{11,1}}{a+b} - \frac{a+c-n_{11,1}}{c+d}\right) - \left(\frac{a}{a+b} - \frac{c}{c+d}\right) \\
&= \left(\frac{1}{a+b} + \frac{1}{c+d}\right)(n_{11,1} - a).
\end{aligned}$$

Therefore, $I(z)$ in Equation (3) can be re-expressed as $I(z)=1$ if $n_{11,1} \le a$ and $I(z)=0$ if $n_{11,1}>a$. Consequently, under the null hypothesis of $n_{10}=n_{01}=0$, the conditional exact p-value can be calculated by

$$\begin{aligned}
p_{a+c,0,0,b+d} &= \sum_{n_{11,1}=\max\{0,a-d\}}^{a+c} I(z) \binom{a+c}{n_{11,1}} \binom{b+d}{a+b-n_{11,1}} \Big/ \binom{n}{a+b} \\
&= \sum_{n_{11,1}=\max\{0,a-d\}}^{a} \binom{a+c}{n_{11,1}} \binom{b+d}{a+b-n_{11,1}} \Big/ \binom{n}{a+b}.
\end{aligned}$$

This is equal to the calculation of the p-value for Fisher's exact test. Therefore, Fisher's exact test can be regarded as a special case of the conditional exact test given here under the null hypothesis of $n_{10}=n_{01}=0$.

Next, we show that Fisher's exact test can be a hypothesis test for the weak causal null hypothesis when monotonicity can be assumed. The null hypothesis of $n_{10}=n_{01}=0$ implies that no subject with type (ii) or (iii) exists, and thus subjects who suffered the event are limited to those with type (i) (i.e., $(Y(1), Y(0))=(1, 1)$), and subjects who did not suffer the event are limited to those with type (iv) (i.e., $(Y(1), Y(0))=(0, 0)$). Therefore, this null hypothesis corresponds to

H_0: $Y(1)=Y(0)$ for all individuals,

which is referred to as the sharp causal null hypothesis [11]. Clearly, whenever the sharp causal null hypothesis holds, the weak causal null hypothesis also holds. However, rejection of the sharp causal null hypothesis does not imply that the weak causal null hypothesis is rejected. This can be explained using the concept of principal stratification as discussed below, and is illustrated using hypothetical data in the following section.

As all subjects must be those with $(Y(1), Y(0))=(1, 1)$ or $(0, 0)$ under the sharp causal null hypothesis, rejection of this hypothesis implies that subjects with $(Y(1), Y(0))=(1, 0)$ or $(0, 1)$ exist. However, even when such subjects are present, if the number of subjects with $(Y(1), Y(0))=(1, 0)$ is equal to the number with $(Y(1), Y(0))=(0, 1)$ (>0), the weak causal null hypothesis still cannot be rejected (i.e., we cannot deny that the true risk difference is zero), because

$$\begin{aligned}
&\Pr(Y(1)=1) - \Pr(Y(0)=1) \\
&= \{\Pr(Y(1)=1, Y(0)=0) + \Pr(Y(1)=1, Y(0)=1)\} \\
&\quad - \{\Pr(Y(1)=0, Y(0)=1) + \Pr(Y(1)=1, Y(0)=1)\} \\
&= \Pr(Y(1)=1, Y(0)=0) - \Pr(Y(1)=0, Y(0)=1) \\
&= 0.
\end{aligned}$$

Consequently, in general, Fisher's exact test cannot be a hypothesis test for the weak causal null hypothesis.

Nevertheless, Fisher's exact test can be a hypothesis test for the weak causal null hypothesis under the following monotonicity assumption [13,14]:

Assumption 1 (Monotonicity):

$Y(0) \le Y(1)$ for all individuals.

This assumption implies that there is no subject with $(Y(1), Y(0))=(0, 1)$. Therefore, under Assumption 1, rejection of the sharp causal null hypothesis implies that there are subjects with $(Y(1), Y(0))=(1, 0)$. Then, the weak causal null hypothesis is also rejected, because

$$\begin{aligned}&\Pr(Y(1)=1)-\Pr(Y(0)=1)\\&=\Pr(Y(1)=1,Y(0)=0)-\Pr(Y(1)=0,Y(0)=1)\\&=\Pr(Y(1)=1,Y(0)=0)\\&>0.\end{aligned}$$

This demonstrates that, under Assumption 1, whenever the sharp causal null hypothesis is rejected, the weak causal null hypothesis is also rejected. Consequently, under the monotonicity assumption, Fisher's exact test is legitimately a hypothesis test for the weak causal null hypothesis.

Barnard's exact test

Barnard's exact test considers the following null hypothesis:

H_0: $\Pr(Y=1 \mid X=1)=\Pr(Y=1 \mid X=0)$,

where $\Pr(Y=1 \mid X=1)$ represents the response proportion for subjects who received the treatment, and $\Pr(Y=1 \mid X=0)$ represents the response proportion for subjects who received the control. Therefore, in general, the null hypothesis for Barnard's exact test is the descriptive null hypothesis to compare two different populations [11], but not the causal null hypothesis to compare the different treatment statuses from one population.

Nevertheless, under randomization, two distributions generated from a single random sample may be the same as those generated by taking two independent random samples [15-17]; i.e., $\Pr(Y(x)=1)=\Pr(Y=1 \mid X=x)$ for $x=0, 1$. If this is true, the following exchangeability assumption [18] must hold:

Assumption 2 (exchangeability):

$\Pr(Y(x)=1 \mid X=1)=\Pr(Y(x)=1 \mid X=0)$ for $x=0, 1$.

This assumption means that for $x=1$, the response proportion for subjects in the treatment group is equal to that if subjects in the control group had received the treatment, and similarly for $x=0$, the response proportion for subjects in the control group is equal to that if subjects in the treatment group had received the control. As a hypothesis test for the causal effect, Barnard's exact test requires the exchangeability assumption. See Greenland [11] for a detailed discussion.

Applying the concept of principal stratification, Assumption 2 implies that when subjects are assigned in a 1:1 ratio by randomization, the numbers of subjects with each type of (i)-(iv) in the treatment group are exactly equal to those in the control group. Although this exact equality may hold at least approximately when the sample size is very large, it may not be true when the sample size is small. For example, by chance, the numbers of subjects with type (i) and (ii) may be greater in the treatment group than in the control group, and instead the numbers of subjects with type (iii) and (iv) may be less in the treatment group than in the control group. Therefore, Assumption 2 may not strictly hold in many randomized trials, and then Barnard's exact test, which requires this assumption, may not adequately test the causal null hypothesis to compare two different treatment statuses from one population.

However, the exact tests given here directly test the weak causal null hypothesis; they do not require that the numbers of subjects with each type of (i)-(iv) in the treatment group are equal to those in the control group when subjects are assigned in a 1:1 ratio by randomization. Therefore, the exact tests do not require Assumption 2. Rather, the exact tests yield the p-value by comparing the risk differences under the null hypothesis generated by violation of Assumption 2 with the risk difference estimated from the observed data. This violation can incidentally be caused as a result of random assignment.

In Table 4, we have summarized the assumptions and the pros and cons of the three exact tests (Fisher, Barnard, and Proposed) for the weak causal null hypothesis.

Extension to Non-Inferiority Trials and Confidence Intervals

Non-inferiority trials

The derived conditional and unconditional exact tests are extended to hypothesis testing for non-inferiority trials below. Hypothesis testing of non-inferiority focuses on the null hypothesis of $\Pr(Y(1)=1) - \Pr(Y(0)=1)=\delta$ rather than $\Pr(Y(1)=1) - \Pr(Y(0)=1)=0$, where δ (> 0) is a small quantity specified in advance. Again, we can express as $\Pr(Y(1)=1)=(n_{11}+n_{10})/n$ and $\Pr(Y(0)=1)=(n_{11}+n_{01})/n$ by using the concept of principal stratification. Therefore, the null hypothesis for non-inferiority, $\Pr(Y(1)=1)-\Pr(Y(0)=1)=\delta$, can be expressed as $n_{10}-n_{01}=\delta n$.

However, when δn is not an integer value, the null hypothesis for the exact tests cannot be prescribed. Therefore, we set the null hypothesis to a maximum integer value satisfying $n_{10}-n_{01} \le \delta n$. Consequently, for non-inferiority trials, the conditional and unconditional exact p-values are calculated by substituting $n_{10}=n_{01}$ in the set of conditions 1 and 2 by $n_{10}-n_{01}=m$, where m is a maximum integer value satisfying $m \le \delta n$.

Confidence intervals

A confidence interval (CI) for a single parameter θ is defined as follows: The interval (L, U) is a 100(1-α)% CI for θ if $\Pr(L \le \theta \le U)=1-\alpha$ [19]. The value of L can be found by seeking a minimum value of the null hypothesis that is not rejected at the significance level $\alpha/2$, and similarly the value of U can be found by seeking a maximum value of the null hypothesis that is not rejected at the significance level $\alpha/2$.

Since the causal risk difference can be expressed as $\Pr(Y(1)=1)-\Pr(Y(0)=1)=(n_{10}-n_{01})/n$, the upper limit of 100(1 - α)% CI for the risk difference, U, linking to the unconditional exact test can be calculated as follows:

	Fisher	Barnard	Proposed
Assumption	Monotonicity	Exchangeability	None
Pros	It is sufficient to test the sharp causal null hypothesis.	It is not a problem to assume exchangeability in randomized trials.	No assumption is required.
Cons	Rejection of the sharp causal null hypothesis does not imply that the causal risk difference is not zero.	Exchangeability may not hold in randomized trials with moderate to small samples.	

Table 4: Assumptions and the pros and cons of the three exact tests (Fisher, Barnard, and Proposed) for the weak causal null hypothesis.

$$U = \sup\left(\frac{n_{10}-n_{01}}{n} : P_U(n_{11}, n_{10}, n_{01}, n_{00}) \geq \frac{\alpha}{2}\right),$$

where $P_U(n_{11}, n_{10}, n_{01}, n_{00})$ is $P_{n_{11},n_{10},n_{01},n_{00}}$ in Equation (1) with set of conditions 1 excluding $n_{10}=n_{01}$, and the lower limit, L, can be calculated as follows:

$$L = \inf\left(\frac{n_{10}-n_{01}}{n} : P_L(n_{11}, n_{10}, n_{01}, n_{00}) \geq \frac{\alpha}{2}\right),$$

where $P_L(n_{11}, n_{10}, n_{01}, n_{00})$ is $P_{n_{11},n_{10},n_{01},n_{00}}$ in Equation (1) with set of conditions 1 excluding $n_{10}=n_{01}$, where the reverse inequality is adopted for the indicator $I(z)$; i.e., $I(z)=1$ if $z \geq 0$ and $I(z)=0$ if $z<0$ with $z := RD_N - RD_O$. To derive the CI linking to the conditional exact test, Equation (2) and set of conditions 1 are replaced by Equation (3) and set of conditions 2.

It is important to note that the upper limit of this CI cannot be larger than the upper bound for the nonparametric bounds [20,21], $\Pr(Y=1, X=1) + \Pr(Y=0, X=0)$, and, likewise, the lower limit cannot be smaller than the lower bound, $-\{\Pr(Y=1, X=0) + \Pr(Y=0, X=1)\}$, even when the sample size is very small. Therefore, the width of CI is always smaller than or equal to 1. This is because

$$\begin{cases} a \leq n_{11} + n_{10} \leq n-b \\ c \leq n_{11} + n_{01} \leq n-d \\ d \leq n_{00} + n_{10} \leq n-c \\ b \leq n_{00} + n_{01} \leq n-a \end{cases},$$

which is derived from the second equation and the fourth inequality in set of conditions 1, corresponds to the nonparametric bounds:

$\Pr(Y=1, X=x) \leq \Pr(Y(x)=1) \leq 1 - \Pr(Y=0, X=x)$ for $x=0, 1$.

Illustration

We illustrate the derived conditional and unconditional exact tests using the data from two clinical trials. The first is a cardiac arrest clinical trial, which is a superiority trial, and the second is an oncology clinical trial, which is a non-inferiority trial. We also show that rejection of the sharp causal null hypothesis does not imply that the weak causal null hypothesis is rejected using the data from a hypothetical clinical trial.

Application to a cardiac arrest clinical trial

Perondi et al. [22] reported a cardiac arrest clinical trial evaluating the next dose of epinephrine to be taken to children suffering cardiac arrest when the initial dose of epinephrine was unsuccessful. In this trial, subjects were randomly assigned in a 1:1 ratio to receive either the same (standard) dose or a higher dose. The endpoint was survival at 24 hours. The results are summarized in Table 5. The risk difference was -0.1765.

The unconditional exact test with $r=1$ yielded the two-sided p-values shown in Figure 1, with several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}). The maximum p-value was 0.0415, which was calculated under $n_{10}=n_{01}=9$. The 95% CI was -0.3382 (-23/68) to -0.0147 (-1/68). The conditional exact test yielded the two-sided p-values shown in Figure 2, with several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}).

	Survival at 24 hours		
Group	**Yes**	**No**	**Total**
Higher dose	1	33	34
Standard dose	7	27	34

Table 5: Results from a cardiac arrest clinical trial.

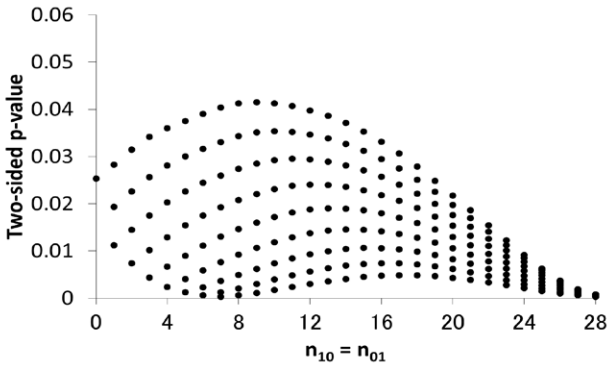

Figure 1: Two-sided p-values for the unconditional exact test under several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}) for the data in Table 5, where the null hypothesis is $n_{10}=n_{01}$.

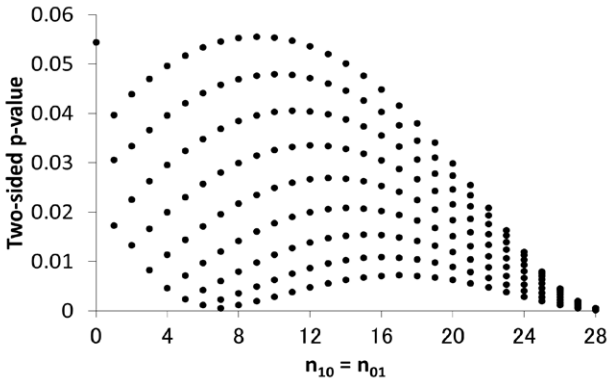

Figure 2: Two-sided p-values for the conditional exact test under several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}) for the data in Table 5, where the null hypothesis is $n_{10}=n_{01}$.

The maximum p-value was 0.0555, which was also calculated under $n_{10}=n_{01}=9$. The 95% CI was -0.3529 (-24/68) to 0 (0/68). We note that, in this example, twice the one-sided p-value was equal to the sum of the one-sided p-value and the opposite one-sided p-value.

In Figure 2, it seems that the p-value under $n_{10}=n_{01}=0$, which corresponds to Fisher's exact test, behaves exceptionally. This is simply because Equation (3) with $n_{10}=n_{01}=0$ is more discrete than Equation (3) with $n_{10}=n_{01}\neq 0$. As discreteness is smaller when the sample size is larger, the extent of the exceptional behavior will be smaller with the larger sample size. Conversely, the p-value under $n_{10}=n_{01}=0$ will be largest for a small sample size, for which violation of the monotonicity assumption (i.e., that at least one subject with $(Y(1), Y(0))=(0, 1)$ exists in the trial) will not be assured.

Application to an oncology clinical trial

Rodary et al. [23] reported an oncology clinical trial, a childhood nephroblastoma study, to demonstrate that pre-operative chemotherapy (new treatment) was not inferior to radiation therapy (standard treatment) in terms of tumor rupture proportions following nephrectomy. The criterion for non-inferiority required that the difference in the proportion of subjects who developed tumor rupture was 0.1 between the chemotherapy (P_C) and radiation (P_R) groups; i.e., the null hypothesis was $P_C - P_R = 0.1$. The subjects were randomly

assigned to either group in a 1:1 ratio. The results are summarized in Table 6. The risk difference was -0.0353.

To apply the conditional and unconditional exact tests, we set the null hypothesis to n_{10}-n_{01}=16 because δn=0.1×164=16.4. The respective unconditional and conditional exact tests yielded the one-sided p-values displayed in Figures 3 and 4, under several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}). The unconditional exact test yielded a maximum p-value of 0.003640 when n_{01}=22 and 95% CI of -0.1280 (-21/164) to 0.0610 (10/164), and the conditional exact test yielded a maximum p-value of 0.003601 when n_{01}=22 and 95% CI of -0.1280 (-21/164) to 0.0610 (10/164).

A hypothetical clinical trial

To demonstrate that rejection of the sharp causal null hypothesis does not imply that the weak causal null hypothesis is rejected, we use the data from a hypothetical randomized clinical trial, shown in Table 7. The risk difference is -0.1000.

	Tumor rupture		
Treatment	**Yes**	**No**	**Total**
Chemotherapy	5	83	88
Radiation	7	69	76

Table 6: Results from an oncology clinical trial.

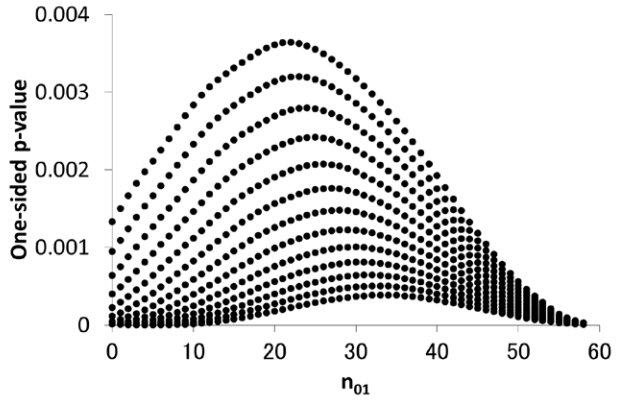

Figure 3: One-sided p-values for the unconditional exact test under several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}) for the data in Table 6, where the null hypothesis is n_{10}-n_{01}=16.

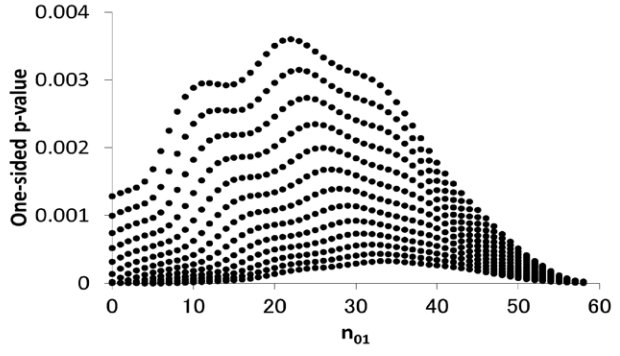

Figure 4: One-sided p-values for the conditional exact test under several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}) for the data in Table 6, where the null hypothesis is n_{10}-n_{01}=16.

	Event		
Group	**Occurred**	**Not occurred**	**Total**
Treatment	1	69	70
Control	8	62	70

Table 7: Results from a hypothetical clinical trial.

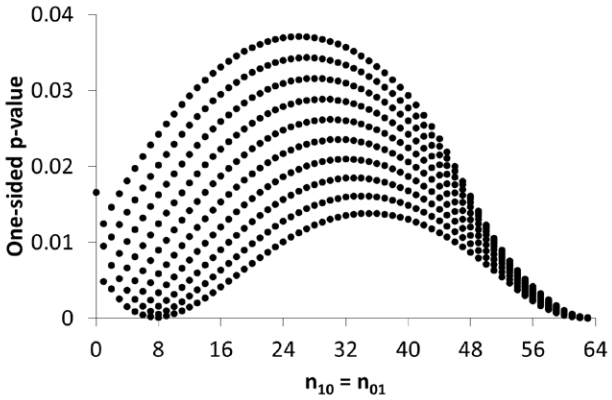

Figure 5: One-sided p-values for the conditional exact test under several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}) for the data in Table 7, where the null hypothesis is n_{10}=n_{01}.

The conditional exact test for the null hypothesis of n_{10}=n_{01} yielded the one-sided p-values shown in Figure 5, with several possible combinations of (n_{11}, n_{10}, n_{01}, n_{00}). The maximum p-value was 0.0371 under n_{10}=n_{01}=26, which corresponds to the p-value for the weak causal null hypothesis. Under n_{10}=n_{01}=0, which corresponds to the p-value for the sharp causal null hypothesis, the p-value was 0.0166. At the significance level of 0.025 (one-sided), the sharp null hypothesis is rejected but the weak causal null hypothesis is not rejected.

As noted in the above cardiac arrest clinical trial, the extent of the exceptional behavior of the p-value under n_{10}=n_{01}=0 will decrease with a larger sample size. This is demonstrated by comparison of Figures 2 and 5. For the larger sample size, there will be more cases in which the sharp null hypothesis is rejected, but the weak causal null hypothesis is not rejected.

Discussion and Conclusion

In this article, we have derived conditional and unconditional exact tests for the weak causal null hypothesis on a binary outcome in randomized trials, using the concept of principal stratification. The derived exact tests have the advantage that they can be extended to non-inferiority trials and to construct CIs in a straightforward manner as a unified approach.

The unconditional exact test will be applied to randomized trials with complete (or equally simple) randomization, and the conditional exact test will be applied to randomized trials with any restriction. However, restricted randomization does not randomly select all a+b subjects of n subjects, and some of them are assigned with dependence on already assigned subjects. Therefore, the conditional exact test may strictly be invalid under restricted randomization. This problem was also pointed out in the context of Fisher's exact test [24].

It might be thought that the exact tests given here should be compared with the existing exact tests for numerical aspects. However, such comparisons would be meaningless, because our new exact tests are hypothesis tests for the weak causal null hypothesis, whereas existing

exact tests are not. It is important to consider which null hypothesis should be tested. In many randomized trials, this will be the weak causal null hypothesis H_0: $\Pr(Y(1)=1)=\Pr(Y(0)=1)$. As an example, let us consider the cardiac arrest clinical trial illustrated in this article. In this trial, researchers set the sample size under the assumption that the 24-hour survival proportion would increase from 20% in the standard dose group to 50% in the higher dose group. This assumption implies that there would certainly be subjects with type 10 ($n_{10} \neq 0$). However, one cannot deny that the 24-hour survival proportion would be higher in the standard dose group than in the higher dose group (i.e., that there would be subjects with type 01 ($n_{01} \neq 0$)). Therefore, at the time of study planning, we should determine to test the weak causal null hypothesis rather than the sharp causal null hypothesis, which corresponds to the null hypothesis of $n_{10}=n_{01}=0$.

The derived exact tests require a numerical search to yield the p-value for the hypothesis testing. The computational effort increases dramatically with the sample size. Therefore, further work is needed to create an efficient algorithm with which the derived exact tests will be feasible. Recently, Rigdon and Hudgens [25] reported a method to construct the CIs applying a similar, but different, approach. Further work will be to compare their CI method with ours.

Acknowledgement

The author thanks the reviewers for helpful comments. This work was supported partially by Grant-in-Aid for Scientific Research (No. 15K00057) from Japan Society for the Promotion of Science.

References

1. Fisher RA (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
2. Fisher RA (1935) The logic of inductive inference. J Roy Stat Soc Ser A 98: 39-82.
3. Barnard GA (1945) A new test for 2×2 tables. Nature 156: 177.
4. Barnard GA (1947) Significance tests for 2 X 2 tables. Biometrika 34: 123-138.
5. Barnard GA (1949) Statistical inference. J Roy Stat Soc Ser B 11: 115-139.
6. Mehrotra DV, Chan IS, Berger RL (2003) A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. Biometrics 59: 441-450.
7. Lydersen S, Fagerland MW, Laake P (2009) Recommended tests for association in 2 x 2 tables. Stat Med 28: 1159-1175.
8. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66: 688-701.
9. Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6: 34-58.
10. Rubin DB (1990) Formal models of statistical inference for causal effects. J Stat Plan Infer 25: 279-292.
11. Greenland S (1992) On the logical justification of conditional tests for two-by-two contingency tables. Am Stat 45: 248-251.
12. Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58: 21-29.
13. Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables (with discussion). J Am Stat Assoc 91: 444-455.
14. Manski CF (1997) Monotone treatment response. Econometrica 65: 1311-1334.
15. Suissa S, Shuster J (1984) Are uniformly most powerful unbiased tests really best? Am Stat 38: 204-206.
16. DAgostino RB, Chase W, Belanger A (1988) The appropriateness of some common procedures for testing equality of two independent binomial proportions. Am Stat 42: 198-202.
17. Little RJA (1989) On testing the equality of two independent binomial proportions. Am Stat 43: 283-288.
18. Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15: 413-419.
19. Cook NR (2005) Confidence intervals and sets. In: Armitage P, Colton T (eds.) Encyclopedia of Biostatistics (2ndedn.) Wiley.
20. Manski CF (1990) Nonparametric bounds on treatment effects. American Economic Review 80: 319-323.
21. Pearl J (1995) Causal inference from indirect experiments. Artif Intell Med 7: 561-582.
22. Perondi MB, Reis AG, Paiva EF, Nadkarni VM, Berg RA (2004) A comparison of high-dose and standard-dose epinephrine in children with cardiac arrest. N Engl J Med 350: 1722-1730.
23. Rodary C, Com-Nougue C, Tournade MF (1989) How to establish equivalence between treatments: a one-sided clinical trial in paediatric oncology. Stat Med 8: 593-598.
24. Routledge R (2005) Fisher's exact test. In: Armitage P, Colton T (eds.) Encyclopedia of Biostatistics (2ndedn). Wiley.
25. Rigdon J, Hudgens MG (2015) Randomization inference for treatment effects on a binary outcome. Stat Med 34: 924-935.