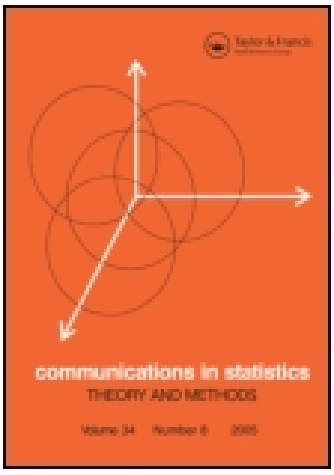


Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

# Exact power and sample size for vaccine efficacy studies

Ivan S.F. Chan [a] & Norman R. Bohidar [a]
[a] Merck Research Laboratories , BL 3-2, West Point, PA, 19486, USA
Published online: 27 Jun 2007.

COMMUN. STATIST.—THEORY METH., 27(6), 1305-1322 (1998)

# EXACT POWER AND SAMPLE SIZE FOR VACCINE EFFICACY STUDIES

Ivan S.F. Chan and Norman R. Bohidar

Merck Research Laboratories, BL 3-2, West Point, PA 19486, USA

*Key Words:* binomial data, conditional test, Poisson distribution, unconditional test



## ABSTRACT

In vaccine efficacy studies the goal is to show that the vaccine reduces the incidence of the disease compared to placebo. In this report we describe two procedures for calculating sample size and power based on exact distributions. In small studies where the disease incidence and the anticipated vaccine efficacy are both high, an unconditional exact procedure is desirable because it guarantees the level of the test and loses little sensitivity. In large studies where the disease incidence is rare, a Poisson approximation to the number of events is reasonable and an exact test is simple to construct conditional on the total number of events. We compare the power and type I error rate of these two exact methods to the method based on the normal approximation for varying disease incidence and sample size.

## 1. INTRODUCTION

The goal of most vaccine efficacy trials is to demonstrate that a new vaccine reduces the incidence of the disease of interest when compared to placebo. Subjects are randomized to receive either the vaccine or a placebo injection, and the response is defined as whether or not the subject develops the disease of interest during the study. The size of a vaccine efficacy trial can range from small scale to mega scale,

depending on the incidence of the disease, the anticipated vaccine efficacy, and the required precision of the study.

In early phase of development of influenza vaccines, for example, a small viral challenge study (see, for example, Fries *et al.* 1993) is often conducted to evaluate the protective efficacy of the vaccine prior to initiating large scale field efficacy trials. Since the attack rate following viral challenge is usually very high, typically a small number of healthy subjects are randomized to receive a new vaccine or placebo, and are subsequently challenged with influenza virus and closely monitored for influenza symptoms. In contrast, a pivotal trial intended to confirm a new vaccine's efficacy usually requires enrolling a large number of subjects, especially when the disease incidence is low. For example, more than a thousand healthy children were enrolled and randomized to receive the vaccine or placebo in a phase III efficacy trial of a new vaccine for prevention of hepatitis A disease, of which the disease incidence rate was predicted to be 3% in the unvaccinated subjects (Werzberger *et al.* 1992). In planning a phase III pivotal trial to evaluate the efficacy of a new vaccine for prevention of the invasive pneumococcal disease in infants, the sample size is estimated to be more than 30,000 subjects because the disease incidence rate is extremely low.

In this report we consider a vaccine efficacy trial where subjects are randomized to receive either a new vaccine or a placebo. Clinicians are primarily interested in whether the vaccine can completely prevent the disease of interest or at least reduce the incidence of the disease. Let $P_1$ and $P_2$ represent the true disease incidence rates among $N_1$ placebo controls and $N_2$ vaccinees, respectively. Let $X$ and $Y$ be the number of cases in the placebo and vaccine groups and are independently distributed as Binomial $(N_1, P_1)$ and Binomial $(N_2, P_2)$ respectively. The vaccine efficacy, denoted by $\pi$, measures the relative reduction of the disease incidence in the vaccine group compared to the placebo group. A simple and commonly used measure of vaccine efficacy is given by (1 - relative risk) for the disease among the vaccinees compared to placebo recipients:

$$\pi = 1 - P_2 / P_1. \tag{1}$$

It is assumed that $P_2 \leq P_1$. The vaccine is 100% efficacious if it prevents the disease completely ($P_2 = 0$); it has no efficacy if $P_2 = P_1$. A test of efficacy focuses on the hypothesis

$$H_0: \pi \le \pi_0 \text{ versus } H_1: \pi > \pi_0 \tag{2}$$

where $\pi_0$ denotes the minimal level of efficacy considered to be acceptable for the new vaccine. While $\pi_0$ could be 0, as is typically the case in establishing therapeutic efficacy of a drug, it is often desirable to demonstrate that the protective efficacy of a new vaccine is significantly greater than some non-zero lower bound ($\pi_0 > 0$) in order to more precisely define the benefit of the vaccine and justify the risk of vaccinating healthy subjects.

Hypothesis (2) is directly related to the hypothesis test on the ratio of two rates. Let $R = P_2 / P_1$, then hypothesis (2) is equivalent to

$$H_0: R \ge 1 - \pi_0 \text{ versus } H_1: R < 1 - \pi_0. \tag{3}$$

A test statistic for hypothesis (2) or (3) has been proposed (Miettinen and Nurminen 1985; Farrington and Manning 1990):

$$Z = \frac{\hat{P}_2 - (1 - \pi_0)\hat{P}_1}{\hat{\sigma}_0 / \sqrt{N_2}} \tag{4}$$

where $\hat{P}_1$ and $\hat{P}_2$ are the observed proportions for the two groups respectively, and

$$\hat{\sigma}_0 = [\tilde{P}_2(1 - \tilde{P}_2) + \frac{1}{c}(1 - \pi_0)^2 \tilde{P}_1(1 - \tilde{P}_1)]^{1/2}, \quad c = N_1 / N_2. \tag{5}$$

Here $\tilde{P}_1$ and $\tilde{P}_2$ are the maximum likelihood estimates of $P_1$ and $P_2$, respectively, constrained under the null hypothesis. Specifically, $\tilde{P}_2$ is the unique solution in (0, 1) of the quadratic equation:

$$(1 + c)x^2 - ax + b = 0 \tag{6}$$

where $a = (1 - \pi_0)(1 + c\hat{P}_1) + c + \hat{P}_2$ and $b = (1 - \pi_0)(c\hat{P}_1 + \hat{P}_2)$. Based on equation (6) one can obtain closed form expressions for $\tilde{P}_1$ and $\tilde{P}_2$ as follows:

$$\tilde{P}_2 = \left\{a - [a^2 - 4b(1 + c)]^{1/2}\right\} / [2(1 + c)], \quad \tilde{P}_1 = \tilde{P}_2 / (1 - \pi_0). \tag{7}$$

Small values of the $Z$ statistic provide evidence against the null hypothesis. The $Z$ statistic is equivalent to the likelihood score statistic (Blackwelder 1993; Nam 1995), and it has a normal approximation in large samples. In analyzing data from small samples, one could use the continuity correction to improve the normal

approximation to the $Z$ statistic given in (4). In this case, one adds $(2N_i)^{-1}$ to $\hat{P}_i$ ($i$ = 1, 2) in the calculation of the $Z$ statistic.

Asymptotic formulas for the power and sample size based on the $Z$ test given in (4) have also been derived. Let $\overline{P}_1$ and $\overline{P}_2$ be the limiting values of $\tilde{P}_1$ and $\tilde{P}_2$ obtained by substituting $P_1$ and $P_2$ in place of $\hat{P}_1$ and $\hat{P}_2$ respectively in calculating the solution given in (7), and substitute them into (5) to get the corresponding limiting value of $\hat{\sigma}_0$, denoted by $\overline{\sigma}_0$. Also let $\sigma_1$ be the value of the expression in (5) when $\tilde{P}_1$ and $\tilde{P}_2$ are replaced with $P_1$ and $P_2$ respectively. Then for testing the null hypothesis $H_0$: $\pi \leq \pi_0$ against a specific alternative $H_1$: $\pi = \pi_1$ ($\pi_1 > \pi_0$), the asymptotic power of an $\alpha$ level $Z$ test, denoted by 1 - $\beta$, is given by

$$1-\beta = \Phi\left\{\frac{1}{\sigma_1}\left[Z_\alpha \overline{\sigma}_0 + N_2^{1/2} P_1(\pi_1 - \pi_0)\right]\right\} \tag{8}$$

where $\Phi$ is the standard Normal distribution function and $Z_\alpha$ is its 100(1 - $\alpha$) percentile. Similarly, the asymptotic sample size expression for the $\alpha$ level $Z$ test with (1 - $\beta$) power is

$$N_2 = \left(Z_\alpha \overline{\sigma}_0 + Z_\beta \sigma_1\right)^2 / \left(P_1(\pi_1 - \pi_0)\right)^2, \quad N_1 = cN_2 \tag{9}$$

where $Z_\beta$ is the 100(1 - $\beta$) percentile of the standard Normal distribution.

Formulas (8) and (9) do not involve the continuity correction. For testing the classical null hypothesis $H_0$: $\pi$ = 0, Casagrande, Pike and Smith (1978) gave an asymptotic formula that incorporates the continuity correction. Other approaches for calculating sample sizes and power are also available for this classical hypothesis (see, for example, Breslow and Day 1987, Chapter 7). Although one could develop asymptotic formulas for the sample size and power incorporating continuity correction for the general case ($H_0$: $\pi \leq \pi_0$), for simplicity, we focus on formulas (8) and (9) for comparion purpose in this report.

When the disease incidence rate and the anticipated vaccine efficacy are both high, a small study may be sufficient to demonstrate vaccine efficacy. The normal approximation to the $Z$ test may not preserve the prespecified type I error rate in small samples, however. Chan (1997) discusses this problem and develops an exact

procedure based on the exact unconditional distribution of the $Z$ statistic. In large scale studies where the disease incidence is rare, the number of cases may be assumed to follow a Poisson assumption, and an exact procedure is simple to construct conditioning on the total number of events. Although the unconditional test could also be used in this setting, it is not practical due to the large computational burden. In this report we compare the power and type I error rate of these two exact methods to the $Z$ test based on the normal approximation for varying disease incidence and sample size.

## 2. EXACT POWER CALCULATION IN SMALL SAMPLES

Chan (1997) proposes an exact unconditional test based on the $Z$ statistic given in (4). This test provides an extension to the tests proposed by Suissa and Shuster (1985) and Haber (1986) to include a non-zero efficacy lower bound in the null hypothesis. The exact test is constructed using the exact unconditional distribution of the test, which consists of all possible outcomes of the two binomial responses given the sample size. Since $X$ and $Y$ are independent, the probability of observing a particular outcome $(i, j)$ is the product of two individual binomial probabilities

$$\Pr(X = i,\ Y = j) = \binom{N_1}{i}\binom{N_2}{j} P_1^i (1-P_1)^{N_1 - i} P_2^j (1-P_2)^{N_2 - j}. \tag{10}$$

This likelihood can be rearranged under the null hypothesis in (2). Let $\pi = \pi_0$ under the null hypothesis. Then we have $P_1 = P_2(1-\pi_0)^{-1}$ and the null likelihood:

$$\begin{aligned}\Pr(X = i,\ Y = j|H_0) = {} & \binom{N_1}{i}\binom{N_2}{j} (1-\pi_0)^{-i} P^{i+j} \\ & \times (1-(1-\pi_0)^{-1}P)^{N_1 - i}(1-P)^{N_2 - j}\end{aligned} \tag{11}$$

where $P = P_2$ is a nuisance parameter.

The exact inference is based on the likelihood (11). For testing the efficacy hypothesis (2), small values of the $Z$ statistic favor the alternative. We use values of the $Z$ statistic to rank the outcomes. Let $Z_{obs}$ be the $Z$ statistic calculated for the observed table. Then the tail (or as extreme region) of the observed table includes all tables for which $Z \le Z_{obs}$. Notice that the likelihood in (11) depends on $P$, which is an unknown nuisance parameter for the exact inference. Since no simple sufficient

**Table I**
**Sample Size Determination Using the Exact Unconditional Test**

| Sample Size Per Group ($N_1 = N_2$) | Critical Value ($Z_c$) | Exact Power (%) | Exact Level (%) |
|---|---|---|---|
| 19 | -2.2808 | 90.3 | 2.48 |
| 20 | -2.2643 | 92.7 | 1.32 |
| 21 | -2.0747 | 95.6 | 2.43 |
| 22 | -2.0067 | 97.4 | 2.37 |
| 23 | -2.2980 | 95.5 | 1.15 |
| 24 | -2.1856 | 97.2 | 1.51 |

Note: The hypothesis $H_0$: $\pi \leq 0.2$ versus $H_1$: $\pi > 0.2$ is tested at the nominal 2.5% level. A true efficacy of 0.8 is assumed under the alternative. The disease incidence rate in the control group is assumed to be 0.8.



statistics for $P$ exist except for the special case where $\pi_0 = 0$, the maximization method proposed by Basu (1977) is used to eliminate the effect of this nuisance parameter. The exact p-value is calculated by maximizing the null likelihood over the domain of $P$, which is $D = [0, 1-\pi_0]$. Specifically, the exact p-value is defined as

$$p = \max_{P \in D} \Pr(Z \leq Z_{obs} \mid \pi_0, P) \tag{12}$$

where the probability is evaluated using the expression in (11).

Given the sample size ($N_1$, $N_2$), one can convert the procedure to find the critical value for an $\alpha$ level exact test of $H_0$: $\pi \leq \pi_0$. This critical value, denoted by $Z_c$, does not depend on any specific value of the nuisance parameter. For a specific alternative $H_1$: $\pi = \pi_1$ ($\pi_1 > \pi_0$), we can calculate $P_2 = (1 - \pi_1)P_1$. Then the power of the $\alpha$ level exact test is the sum of the probabilities of those tables for which the $Z$ statistics are less or equal to $Z_c$:

$$1 - \beta = \Pr(Z \leq Z_c \mid P_1, P_2) \tag{13}$$

where the probability is evaluated using the expression in (10). Similarly, one can use the procedure to find the minimum sample size required for a study to achieve a pre-planned power.

As an example to illustrate the sample size calculation, suppose the disease incidence rate is 80% in the placebo control group, and the new vaccine is expected

to have 80% efficacy. Suppose the study is required to have at least 95% power to show that the vaccine efficacy exceeds the lower bound of 20% using a 2.5% level exact test. Assuming equal sample sizes in both groups, Table I gives the critical value, power and level of the exact unconditional test for a range of sample sizes. Since the exact distribution is discrete, as is the case in Fisher's exact test, the nominal level cannot be fully achieved. As a result, the power function is not strictly monotone and may have some little dips when the sample size increases. By examining a range of sample sizes we find the minimal sample size required is 21 per group to achieve the 95% power. The associated true level is 2.43%. This sample size is close to the asymptotic sample size (19 per group) obtained from formula (9) for the Z test.



## 3. EXACT POWER CALCULATION IN LARGE SAMPLES UNDER POISSON ASSUMPTION

In assessing a new vaccine's efficacy against a rare disease such as polio or the invasive pneumococcal disease, a large scale clinical trial is often required in order to precisely estimate the vaccine efficacy. When the incidence rate of the disease is very low and the sample size is very large, it is possible to construct an exact conditional test assuming the number of disease cases occurring in the study follow a Poisson distribution (Gail 1974; Breslow and Day 1987, Section 7.3; Guess *et al* 1987). Sample sizes and power calculations for the classical hypothesis $H_0$: $\pi = 0$ have been extensively studied by Gail (1974), and Breslow and Day (1987, Section 7.3). Bohidar and Chan (1996) make a simple extension to include null hypotheses with non-zero lower bounds. We briefly describe the procedure in the following.

Suppose $N_1$ and $N_2$ are sufficiently large, and $P_1$ and $P_2$ are sufficiently small so that $N_1 P_1 \to \lambda_1$ and $N_2 P_2 \to \lambda_2$, then $X$ and $Y$ are approximately distributed as independent Poisson random variables with parameters $\lambda_1$ and $\lambda_2$, respectively. Let $T$ be the total number of cases observed in the study. Then conditional on $T$, the number of cases in the vaccine group ($Y$) is distributed as Binomial ($T$, $\theta$) where

$$\theta = \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{N_2 P_2}{N_1 P_1 + N_2 P_2} = \frac{1-\pi}{1+c-\pi}, \quad c = N_1 / N_2. \tag{14}$$

Since $\theta$ is decreasing in $\pi$, the efficacy hypothesis in (2) is equivalent to

$$H_0: \theta \geq \theta_0 \text{ versus } H_1: \theta < \theta_0 \tag{15}$$

where $\theta_0 = (1-\pi_0)/(1+c-\pi_0)$. Suppose $Y_{obs}$ is the number of cases observed in the vaccine group, then the exact conditional p-value is

$$p = \Pr[Y \leq Y_{obs} | Y \sim \text{Binomial}(T,\theta_0)] = \sum_{k=0}^{Y_{obs}} \binom{T}{k} \theta_0^k (1-\theta_0)^{T-k}. \tag{16}$$

To calculate the power for the efficacy hypothesis in (2) against a specific alternative where $\pi = \pi_1 (> \pi_0)$, we first compute $\theta_1 = (1-\pi_1)/(1+c-\pi_1)$, and the critical value, denoted by $Y_c$, at the $\alpha$ level. Then the power is given by

$$1-\beta = \Pr[Y \leq Y_c | Y \sim \text{Binomial}(T,\theta_1)] = \sum_{k=0}^{Y_c} \binom{T}{k} \theta_1^k (1-\theta_1)^{T-k}. \tag{17}$$

This conditional approach can also be used in planning a study. That is, the study can be designed to accrue a fixed number of events ($T$) rather than to run for a fixed duration. Since the unconditional expected value of $T$ is $(N_1P_1 + N_2P_2)$, one can estimate the expected number of subjects required for the study using the incidence rate of the disease in the placebo group ($P_1$) and the expected vaccine efficacy under the alternative hypothesis ($\pi_1$) using the following formula

$$N_2 \approx T/[(c+1-\pi_1)P_1]. \tag{18}$$

One important feature of the fixed number of events design is that the power of the study depends on the disease incidence rates only through their ratio ($P_2/P_1$) or the anticipated efficacy ($\pi_1$). With this fixed information approach, one can avoid the situation where the anticipated power is not achieved because the number of cases that occurs by the study completion is fewer than expected. This is in contrast to a fixed duration trial where the actual power may be lower than anticipated if the disease incidence rate in the placebo group turns out to be lower than expected. However, if the true vaccine efficacy is much higher than expected, it may take many more subjects or much longer time than anticipated to observe the total $T$ cases in a fixed number of events trial. As a result, it may be useful to plan an interim analysis to help potentially avoid this situation.

Most large scale efficacy studies will also have a long study duration, and thus differential amount of follow-up time may occur in the two groups due to a number of reasons. This will cause a problem for the $Z$ test because it assumes equal amount of follow-up in the two groups. However, the Poisson model is readily extendable to

**Table II**
**Sample Size Determination Using the Exact Conditional Test Based on Poisson Assumption**

| Total Number of Cases ($T$) | Critical Value ($Y_c$) | Exact Power (%) | Exact Level (%) |
|---|---|---|---|
| 33 | 8 | 91.4 | 1.36 |
| 34 | 9 | 95.4 | 2.44 |
| 35 | 9 | 94.5 | 1.79 |
| 36 | 9 | 93.4 | 1.30 |
| 37 | 10 | 96.5 | 2.28 |
| 38 | 10 | 95.8 | 1.68 |
| 39 | 10 | 95.0 | 1.23 |
| 40 | 11 | 97.4 | 2.11 |

Note: The hypothesis $H_0$: $\pi \leq 0.2$ versus $H_1$: $\pi > 0.2$ is tested at the nominal 2.5% level. A true efficacy of 0.8 is assumed under the alternative.

this situation. The only modification required for the exact conditional test is on the binomial parameter $\theta$ given in (14). When there is differential follow-up time, one can modify the constant $c$ to be the ratio of follow-up time instead of sample sizes between the two groups.

To illustrate the sample size determination using the exact conditional method, consider an example of testing $H_0$: $\pi \leq 0.2$ against $H_1$: $\pi > 0.2$. Suppose we design the study to have a 95% power at the 2.5% level assuming a true vaccine efficacy of 0.8 for a rare disease. Table II lists the power and level of the exact test for a range of total number of cases. The power reaches 95.4% when the total number of cases is 34. Due to the discrete nature of the exact distribution, however, the power drops slightly below 95% when the total number of cases increase to 35 and 36. The power goes back and stays above 95% when the total number of cases is 37 or larger. In the planning stage, we recommend taking a conservative approach and consider $T$ = 37 as the minimum total number of cases required for the study. Thus, the study is ensured to have at least 95% power even if a couple more cases are observed in the study. Assuming equal sample sizes in both groups and a disease incidence rate of 0.6% in the placebo control group, the total number of subjects needed for enrollment is projected to be 10,278 using the formula in (18). For comparison, the sample size based on the asymptotic formula for the Z test is 10,838.

## 4. COMPARISONS OF THE EXACT PROCEDURES AND THE NORMAL APPROXIMATION TEST

In this section we compare the power and size of the two exact tests to the $Z$ test based on the normal approximation for varying disease incidence and sample size. We show that these two exact tests are nearly as sensitive as the normal approximation test.

### 4.1 Comparisons in Small Samples

In the first set of comparisons we consider testing the hypothesis $H_0$: $\pi \leq 0.4$ versus $H_1$: $\pi > 0.4$ at the nominal 5% level. In the power calculation, we assume a disease incidence rate ($P_1$) of 90% in the placebo control group, a true efficacy ($\pi_1$) of 0.8, and equal sample sizes in both groups. Table III gives the size and power of the exact unconditional test proposed by Chan (1997) and the $Z$ test given in (4) based on normal approximation for various sample sizes. The true size and power of the $Z$ test is calculated by enumerating all possible combinations of responses given the sample size, true efficacy, and the disease incidence rate in the control group. It can be seen that the true size of the $Z$ test based on the normal approximation is typically higher than that of the exact test and in some cases well exceeds the nominal level of 5%. In contrast, the size of the exact test never exceeds 5%. Note that by construction, the size of the exact test is the maximum over the domain of the nuisance parameter. As a result, it represents the worst case scenario and does not depend on the incidence rate in the control group. On the other hand, the true size of the $Z$ test changes when the incidence rate in the control group ($P_1$) changes. The fourth column of Table III gives the maximum type I error rate of the $Z$ test over all possible values of $P_1$ and shows that it is almost always higher than the nominal level. Thus, the use of the $Z$ test in this range of sample size may be too liberal.

It is interesting to observe that the asymptotic power of the $Z$ test calculated based on the normal approximation given in (8) estimates the true power very well even when the sample size is small. The exact test has less power than the $Z$ test when the sample size per group is 12 or less. Of course this can be partially explained by the fact that the $Z$ test is too liberal with these small sample sizes. The power of the exact test increases and gets very close to the power of the $Z$ test when the sample size increases. In fact, it appears that the exact test can be nearly as sensitive as the $Z$ test when the study is designed with reasonable power (~80% or

**Table III**
**The Size and Power of the Exact Unconditional Test and the Normal Z Test at the 5% Level in Small Samples**

| | Size of Test (%) | | | Power of Test (%) | | |
|---|---|---|---|---|---|---|
| | | Z Test Based on Normal Approximation | | | Z Test Based on Normal Approximation | |
| Sample Size Per Group ($N_1 = N_2$) | Exact Unconditional Test | True | Maximum | Exact Unconditional Test | True | Asymptotic |
| 5 | 3.64 | 9.03 | 9.04 | 34.1 | 58.1 | 39.6 |
| 6 | 4.83 | 4.48 | 4.83 | 51.2 | 51.2 | 46.7 |
| 7 | 4.43 | 9.52 | 9.70 | 43.2 | 69.5 | 53.4 |
| 8 | 4.95 | 5.06 | 5.11 | 47.7 | 61.5 | 59.5 |
| 9 | 3.64 | 4.56 | 5.83 | 59.4 | 70.7 | 65.0 |
| 10 | 1.92 | 5.34 | 6.90 | 52.2 | 71.6 | 69.9 |
| 11 | 4.74 | 5.25 | 6.13 | 71.8 | 79.3 | 74.3 |
| 12 | 3.53 | 5.42 | 5.73 | 73.8 | 79.4 | 78.2 |
| 13 | 4.69 | 3.59 | 5.86 | 79.8 | 80.6 | 81.6 |
| 14 | 3.74 | 5.39 | 5.81 | 80.9 | 85.1 | 84.5 |
| 15 | 4.43 | 3.83 | 5.47 | 85.5 | 86.6 | 87.0 |
| 16 | 3.83 | 5.30 | 5.79 | 86.0 | 89.3 | 89.1 |
| 17 | 4.50 | 5.84 | 5.95 | 89.6 | 91.1 | 91.0 |
| 18 | 3.87 | 5.19 | 5.71 | 89.7 | 92.7 | 92.5 |
| 19 | 4.75 | 5.76 | 5.79 | 92.5 | 93.3 | 93.8 |
| 20 | 3.69 | 5.05 | 5.59 | 91.2 | 94.4 | 94.9 |
| 21 | 4.61 | 5.62 | 5.85 | 94.5 | 94.5 | 95.8 |
| 22 | 3.63 | 4.92 | 5.42 | 93.4 | 96.0 | 96.6 |
| 23 | 4.65 | 5.43 | 5.61 | 96.0 | 96.3 | 97.2 |
| 24 | 3.68 | 4.77 | 5.80 | 95.1 | 97.1 | 97.7 |
| 25 | 4.27 | 5.41 | 5.49 | 97.0 | 97.8 | 98.1 |

Note: The hypothesis tested is $H_0$: $\pi \leq 0.4$ versus $H_1$: $\pi > 0.4$. A true efficacy of 0.8 is assumed under the alternative, and an incidence rate of 0.9 is assumed for the control group.

**Table IV**
**Sample Sizes Required for the Exact Unconditional Test and the Normal $Z$ Test at the 2.5% Level with 90% Power**

| | | | Exact Unconditional Test | | $Z$ Test Based on Normal Approximation | |
|---|---|---|---|---|---|---|
| Incidence Rate in Control Group ($P_1$) | True Efficacy ($\pi_1$) | Efficacy Lower Bound ($\pi_0$) | Sample Size Per Group ($N_1 = N_2$) | True Size (%) | Asymptotic Sample Size Per Group ($N_1 = N_2$) | True Size (%) |
| .9 | .8 | 0 | 9 | 1.64 | 8 | 0.36 |
| .9 | .8 | .4 | 22 | 2.25 | 21 | 2.98 |
| .9 | .8 | .6 | 68 | 2.47 | 69 | 2.23 |
| .9 | .4 | 0 | 31 | 2.40 | 31 | 2.76 |
| .9 | .4 | .1 | 53 | 1.84 | 49 | 2.90 |
| .7 | .7 | 0 | 21 | 2.48 | 20 | 2.79 |
| .7 | .7 | .35 | 55 | 2.48 | 54 | 2.37 |
| .7 | .5 | 0 | 43 | 2.48 | 41 | 2.47 |
| .7 | .5 | .1 | 62 | 2.48 | 58 | 3.07 |
| .7 | .5 | .2 | 97 | 2.47 | 93 | 2.52 |
| .5 | .8 | 0 | 26 | 2.15 | 26 | 2.25 |
| .5 | .8 | .4 | 62 | 2.30 | 61 | 2.33 |
| .5 | .5 | 0 | 78 | 2.46 | 77 | 2.26 |
| .5 | .5 | .1 | 110 | 2.44 | 108 | 2.53 |

higher). This finding is consistent with our observations in many other examples not reported here.

To further assess the efficiency of the exact test, we compare the minimum sample size required for the two tests to have at least 90% power at the 2.5% nominal level. Table IV presents the sample size requirement for the two tests under various configurations of the hypothesis and varying incidence rates in the control group. The sample size for the $Z$ test is calculated using the asymptotic formula in (9). It is clear from the table that the sample size for the exact test compares very favorably to that for the $Z$ test under all different scenarios of hypothesized efficacy lower bound, true efficacy, and incidence rates in the control group. Again, the true

size of the $Z$ test may exceed the nominal level. Although the true power of the $Z$ test is not reported here, it always falls in the range from 90% to 93%. The exact test achieves very similar power in all cases, and the biggest difference in power between the two methods is 2 percentage points. These results again demonstrate that the exact test not only controls the type I error rate, but also is efficient when compared to the $Z$ test.

### 4.2 Comparisons in Large Samples

We compare the power and size of the exact conditional test based on Poisson assumption to the $Z$ test for the efficacy hypothesis $H_0$: $\pi \leq \pi_0$ versus $H_1$: $\pi > \pi_0$. Both tests are performed at the nominal 2.5% level with the assumption of equal sample sizes in both groups. The parameters that affect the power and size of the test are the incidence rate in the control group ($P_1$), the efficacy lower bound ($\pi_0$), the true efficacy in the alternative hypothesis ($\pi_1$), and the sample size. Table V summarizes some representative comparisons of the power and size of the two tests under a variety of settings of these parameters. The asymptotic power of the Z test is obtained using formula (9), and the true power and size are calculated by enumerating all possible combinations of the two binomial responses. For the exact test, we calculate the total number of cases expected in the two groups ($T$) using $P_1$ and $\pi_1$; then we use formulas (16) and (17) to compute the expected size and expected power of the exact test.

The results in Table V suggest that the size of the exact conditional test is almost always less than the size of the Z test, suggesting that the exact test is a little bit conservative. The true size of the $Z$ test sometimes exceeds the nominal level. Other factors held constant, the conservatism decreases when the expected total number of cases increases. When the incidence rate is high ($\geq$ 0.3), the exact test has substantially less power than the Z test; and the lack of power becomes more serious when incidence rates increase or sample sizes decrease. On the other hand, the Poisson approximation works very well for very small incidence rates ($\leq$ 0.01). In these cases, the power of the exact test is very close to that of the $Z$ test for almost all configurations of $\pi_0$ and $\pi_1$. As the incidence rate increases to up to 0.1, the power of the exact test tends to be slightly lower than the $Z$ test. However, if the sample size is big enough to have reasonable power ($\geq$ 80%), the exact test can be as good as the $Z$ test. As in the small samples, the power of the $Z$ test calculated using the asymptotic formula provides very accurate estimate of the true power.

**Table V**
**The Size and Power of the Exact Conditional Test and the Normal $Z$ Test at the 2.5% Level in Large Samples**

| Incidence Rate in Placebo Group ($P_1$) | True Efficacy ($\pi_1$) | Efficacy Lower Bound ($\pi_0$) | Total Sample Size | Expected Total Number of Cases ($T$) | Exact Conditional Test | | $Z$ Test Based on Normal Approximation | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Expected Power (%) | True Size (%) | Asymptotic Power (%) | True Power (%) | True Size (%) |
| .3 | .6 | 0 | 200 | 42 | 80.5 | 2.18 | 88.4 | 89.3 | 2.55 |
| .3 | .6 | .4 | 1,400 | 294 | 88.9 | 2.09 | 95.0 | 95.0 | 2.42 |
| .1 | .8 | .2 | 400 | 24 | 80.0 | 1.46 | 82.4 | 84.9 | 2.41 |
| .1 | .8 | .4 | 800 | 48 | 90.8 | 2.35 | 89.6 | 90.9 | 2.34 |
| .1 | .6 | .4 | 2,000 | 140 | 54.3 | 1.68 | 62.2 | 62.4 | 2.41 |
| .1 | .6 | .4 | 4,000 | 280 | 86.9 | 1.99 | 90.2 | 90.4 | 2.42 |
| .05 | .8 | .2 | 1,200 | 36 | 93.5 | 1.30 | 94.0 | 95.5 | 2.35 |
| .05 | .4 | 0 | 2,000 | 80 | 55.0 | 1.65 | 62.7 | 63.1 | 2.49 |
| .05 | .4 | 0 | 4,000 | 160 | 88.9 | 2.39 | 89.8 | 90.2 | 2.50 |
| .03 | .8 | 0 | 1,000 | 18 | 83.2 | 1.54 | 81.5 | 85.2 | 2.47 |
| .03 | .4 | 0 | 4,000 | 96 | 62.7 | 1.58 | 69.8 | 70.4 | 2.50 |
| .03 | .4 | .2 | 18,000 | 432 | 82.8 | 2.32 | 84.0 | 84.1 | 2.48 |
| .02 | .75 | 0 | 800 | 10 | 37.6 | 1.07 | 48.0 | 47.0 | 2.32 |
| .02 | .75 | 0 | 2,000 | 25 | 89.1 | 2.16 | 85.6 | 88.9 | 2.58 |
| .02 | .75 | .25 | 4,000 | 50 | 93.9 | 2.21 | 92.3 | 93.6 | 2.36 |
| .02 | .4 | 0 | 4,000 | 64 | 45.3 | 1.64 | 52.2 | 52.4 | 2.48 |
| .02 | .4 | 0 | 14,000 | 226 | 95.9 | 1.95 | 96.5 | 96.7 | 2.49 |
| .01 | .8 | 0 | 2,000 | 12 | 67.7 | 1.93 | 63.9 | 65.9 | 2.54 |
| .01 | .8 | .2 | 6,000 | 36 | 93.5 | 1.30 | 93.3 | 95.3 | 2.41 |
| .01 | .8 | .4 | 8,000 | 48 | 90.8 | 2.35 | 87.9 | 89.6 | 2.30 |
| .01 | .5 | 0 | 4,000 | 30 | 43.2 | 2.14 | 44.9 | 45.8 | 2.42 |
| .01 | .5 | 0 | 12,000 | 90 | 89.0 | 2.23 | 88.8 | 89.5 | 2.49 |
| .005 | .8 | 0 | 4,000 | 12 | 67.7 | 1.93 | 63.8 | 65.9 | 2.57 |
| .005 | .8 | 0 | 10,000 | 30 | 98.0 | 2.14 | 95.5 | 97.6 | 2.44 |

**Table VI**
**Total Sample Sizes Required for the Exact Conditional Test and the Normal Z Test in Large Samples**

| Incidence Rate in Placebo Group ($P_1$) | True Efficacy ($\pi_1$) | Efficacy Lower Bound ($\pi_0$) | Nominal Significance Level (%) | Required Power (%) | Exact Conditional Test: Total Number of Cases | Exact Conditional Test: Expected Sample Size | $Z$ Test via Normal Approximation: Asymptotic Sample Size | Efficiency ($Z$ Test/Exact) × 100% |
|---|---|---|---|---|---|---|---|---|
| .05 | .8 | 0 | 2.5 | 80 | 17 | 568 | 570 | 100.4 |
| .05 | .8 | 0 | 2.5 | 90 | 23 | 768 | 766 | 99.7 |
| .05 | .8 | 0 | 2.5 | 95 | 28 | 934 | 940 | 100.6 |
| .05 | .8 | .4 | 2.5 | 95 | 61 | 2,034 | 2,32 | 99.9 |
| .05 | .8 | 0 | 5.0 | 90 | 32 | 1,068 | 1,026 | 96.1 |
| .01 | .9 | 0 | 2.5 | 90 | 17 | 3,092 | 3,506 | 113.4 |
| .01 | .9 | .4 | 2.5 | 85 | 21 | 3,818 | 4,192 | 109.8 |
| .01 | .9 | .4 | 5.0 | 95 | 26 | 4,728 | 4,782 | 101.1 |
| .01 | .5 | 0 | 2.5 | 85 | 86 | 11,468 | 10,692 | 93.2 |
| .01 | .5 | 0 | 2.5 | 95 | 119 | 15,868 | 15,472 | 97.5 |
| .005 | .8 | .2 | 2.5 | 90 | 32 | 10,666 | 10,610 | 99.5 |
| .005 | .8 | .2 | 5.0 | 80 | 21 | 7,000 | 6,290 | 90.0 |
| .003 | .75 | .25 | 2.5 | 85 | 42 | 22,400 | 21,346 | 95.3 |
| .003 | .75 | .25 | 2.5 | 95 | 56 | 29,866 | 30,412 | 101.8 |

We also compare the performance of the two tests in terms of sample size requirement. For the exact test, we first calculate the minimum requirement of the total number of cases in the two groups ($T$) required. Then we use the incidence rates in the control and vaccine groups to project the number of subjects needed in the study. Table VI lists the sample size requirement for the two tests under various settings. In general, the sample sizes given by the two tests are close with most differences within 5%. The efficiency of the exact test seems to be pretty robust to the incidence rates and the true efficacy as long as the study is adequately powered.

## 5. DISCUSSION

In this report we consider the power and sample size calculations of two exact procedures for testing efficacy hypotheses. We also compare the performance of these tests to the method based on the normal approximation. Since the efficacy is directly related to the relative risk, as indicated by the equivalence of hypotheses (2) and (3), these methods can also be applied to testing non-unity relative risk between a new vaccine and a standard vaccine. In designing a study, it is very important that the minimum efficacy ($\pi_0$) or maximum risk ratio be specified in the protocol prior to the start of the study and the power be planned accordingly.

The exact unconditional test proposed by Chan (1997) uses the unconditional distribution of the test statistic. The test controls the type I error rate and can be as sensitive as the $Z$ test. This exact test works well in small to moderate samples. However, the computational effort involved increases drastically with the sample size since the exact procedure requires numerical search for the maximum rejection probability. As a result, the exact unconditional test may not be practical in studies where a very large number of subjects are involved. In searching over the domain of the nuisance parameter, we use a uniform search with 100 grid points in our power calculations. An increase of the number of grid points to 1,000 may slighly improve the precision but substantially increase the computational time. The computer programs for the exact unconditional test and the exact conditional test are available from the first author upon request.

When using the exact unconditional test in analyses, one can sometimes improve the sensitivity of the test using the method proposed by Berger and Boos (1994). According to their method, one can first construct a $100(1-\Upsilon)\%$ confidence interval on the nuisance parameter, denoted by $C_\Upsilon$, and then search for the maximum tail probability within the confidence interval. The exact p-value will be defined as $p_\gamma = \max_{P \in C_\Upsilon} \Pr(Z \leq Z_{obs} | \pi_0, P) + \gamma$ where $\gamma$ serves as a penalty for restricting the search for the maximum tail probability to the confidence set. The choice of $\gamma$ is usually very small ($\leq 0.0001$) to avoid too much penalty on the p-value. This method has been implemented in the commercial software StatXact (1995) for similar problems.

In large scale studies of rare disease, the exact conditional test based on the Poisson assumption compares very favorably to the $Z$ test based on the normal approximation. The exact test is conditional on the total number of the disease cases. It only focuses on the split of cases between the vaccine and the placebo group and does not depend on the incidence rate in the placebo group. In addition to its simplicity, the exact test can be implemented in the fixed number of events design to ensure the power of the study. It can also be easily extended to handle differential follow-up time between groups.

## ACKNOWLEDGEMENTS

The authors would like to thank Holly Matthews and Joseph Heyse for their encouragement and support for this research, and the guest editor and referees for their helpful suggestions.

## BIBLIOGRAPHY

Berger, R.L. and Boos, D.D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *Journal of American Statistical Association*, 89, 1012-1016.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of American Statistical Association*, 72, 355.

Blackwelder, W.C. (1993). Sample size and power for prospective analysis of relative risk. *Statistics in Medicine*, 12, 691-698.

Bohidar, N. and Chan, I.S.F. (1996). An evaluation of exact power calculations for vaccine efficacy trials. Abstract, International Biometric Conference, Amsterdam, Netherlands.

Breslow. N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research*, Volume II - The Design and Analysis of Cohort Studies. Oxford University Press, Oxford, UK.

Casagrande, J.T., Pike, M.C., and Smith, P.G. (1978). An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, 34, 483-486.

Chan, I.S.F. (1997). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. To appear in *Statistics in Medicine*.

Farrington, C.P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9, 1447-1454.

Fries, L.F. *et al.* (1993). Safety and immunogenicity of a recombinant protein influenza A vaccine in adult human volunteers and protective efficacy against wild-type H1N1 virus challenge. *Journal of Infectious Diseases*, 167, 593-601.

Gail, M. (1974). Power computations for designing comparative Poisson trials. *Biometrics*, 30, 231-237.

Haber, M. (1986). An exact unconditional test for the 2 × 2 comparative trials. *Psychological Bulletin*, 99, 129-132.

Guess, H.A., Lydick, E.G., Small, R.D. and Miller, L.P. (1987). Epidemiologic programs for computers and calculators - Exact binomial confidence intervals for the relative risk in follow-up studies with sparsely stratified incidence density data. *American Journal of Epidemiology*, 125, 340-347.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4, 213-226.

Nam, J.M. (1995). Sample size determination in stratified trials to establish the equivalence of two treatments. *Statistics in Medicine*, 14, 2037-2049.

Suissa, S. and Shuster, J.J. (1985). Exact unconditional sample sizes for the 2 × 2 binomial trial. *Journal of Royal Statistical Society*, Series A, 148, 317-327.

*StatXact 3* (1995). Cytel Software Corporation, MA.

Werzberger, A. *et al.* (1992). A controlled trial of a formalin-inactivated hepatitis A vaccine in healthy children. *New England Journal of Medicine*, 327, 453-457.