Chop-lump Tests for Vaccine Trials

Dean Follmann, Michael Fay, & Michael Proschan[1]
September 4, 2007

ABSTRACT

This paper proposes new tests to compare the vaccine and placebo groups in randomized vaccine trials when a fraction of volunteers become infected. A simple approach that is consistent with the intent-to-treat principle is to assign a score, say W, equal to 0 for the uninfecteds and some disease severity $X > 0$ for the infecteds. One can then test the equality of this lumpy distribution of W between the two groups. This Burden of Illness (BOI) test was introduced by Chang, Guess, & Heyse (1994). If infections are rare, the massive number of 0s in each group tends to dilute the vaccine effect and this test can have poor power, particularly if the Xs are not close to zero. Comparing X in just the infecteds is no longer a comparison of randomized groups and can produce misleading conclusions. Hudgens, Hoering, & Self (2003) introduced tests of the equality of X in a subgroup—the principle stratum of those who are doomed to be infected no matter what. This can be more powerful than the BOI approach, but requires the assumption that the vaccine cannot directly or indirectly cause infections in anyone. We suggest new "chop-lump" tests that can be more powerful than the BOI tests in certain situations, and do not require the assumption that the vaccine harms no one. The basic idea is to toss out an equal proportion of zeros from both groups and then perform a test on the remaining Ws which are mostly > 0. A permutation approach provides a null distribution. This type of test can be viewed as a BOI test in the principal stratum of the non-immune under certain assumptions and is always a good test when the difference between two distributions is concentrated in one tail. We show that under local alternatives, the chop-lump Wilcoxon test is always more powerful than the usual Wilcoxon test asymptotically, provided the underlying infection rates are the same. We also identify the crucial role of the "gap" between 0 and the average X on power for the t-tests. The chop-lump tests are compared to established tests via simulation for planned HIV and malaria vaccine trials. A re-analysis of the first phase III HIV vaccine trial is used to illustrate the method.

KEY WORDS: Lachenbruch's tests Permutation; Principal Stratification; Rank Tests; Semi-continuous data.

[1]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, 6700B Rockledge Drive MSC 7609, Bethesda, MD 20892. (E-mail: *dfollmann@niaid.nih.gov*)

1 Introduction

Definitive evaluation of candidate vaccines requires a properly designed randomized clinical trial. For vaccines intended only to reduce the disease burden of the infected, but not prevent infection, analysis can be complicated, especially for trials where infection is relatively rare. The simplest approach is to assign a score of zero to the uninfecteds and use some positive measure of disease severity, say X in the infecteds. The distributions of the resulting lumpy composite random variable, say W can be compared between all randomized participants. Chang, Guess, & Heyse (1994) suggested comparing the two groups using the differences in means of W, calling W the Burden of Illness. This is appealing as all randomized participants are explicitly included in the analysis and the intent to treat principle is followed. However, it can have poor power in some settings.

A more powerful approach can be to compare the distribution of X in the infecteds. However, the vaccine and placebo infecteds are not ensured to be comparable by randomization, except under the null hypothesis that the vaccine has no effect on acquisition of infection. If the vaccine does nothing but cause some infections with low disease severity, a test in the infecteds might conclude the vaccine is beneficial when in fact it is harmful. In general a vaccine might "cause" infections by a variety of mechanisms. For example some vaccines are actually live but weakened pathogens. If the weakened pathogen is used in a weakened host, disease producing infection might result. This type of problem was the reason the US switched from an attenuated polio vaccine to a killed vaccine (Alexander et al 2004). Another cause is enhancing antibodies, vaccine induced proteins of the immune system that "enhance" rather than reduce the chance a disease producing infection occurs (Burke 1992). Additionally, vaccines might induce an auto-immune reaction which could hamper the ability of the immune system to fight infection or disease. Finally, a vaccine might encourage risky behavior that results in more exposure to pathogens, thus leading to

an increased probability of infection. Thus it is appealing to have methods of analysis that avoid the assumption that the vaccine does not cause infections.

Gilbert, Bosch, & Hudgens (2003) introduced a class of tests for the distribution of X (i.e. post-infection viral load) that addressed selection bias. They applied the thinking of Frangakis and Rubin (2000) and postulated a principal stratum of volunteers doomed to become infected whether randomized to placebo or vaccine. If a trial could be done in the doomed, all would be infected and there would be no concern about selection bias. In practice, the doomed cannot be identified, and any trial will include the doomed and others. GBH construct tests by essentially culling out placebo infecteds who are unlikely to be from the doomed stratum according to a model indexed by a single parameter reflecting the amount of selection bias. The remaining placebo infecteds are then compared to the vaccine infecteds. Shepherd et al (2006) generalize this thinking to the regression setting and extend to studies other than vaccine trials. Hudgens, Hoering, & Self (2003) focused on extreme versions of the test, in particular, where the worst (i.e. largest) viral loads are eliminated from the placebo group, thus maximally penalizing the vaccine within this class of tests. But these approaches must assume that the vaccine cannot cause infections.

Lachenbruch proposed an omnibus procedure by combining a test of the proportion infected using all participants with a t-test among the infecteds. Each test statistic is squared and the two are added. This has a chi-squared distribution with 2 degrees of freedom asymptotically. However, if the vaccine is expected to have little effect on acquisition, this approach may be less powerful than a procedure that focuses on the Xs. Tu & Zhou (1999) proposed a parametric approach modelling the Ws with a mixture of point mass at zero and a log-normal distribution. A likelihood ratio test was derived that tested equality of the 3 parameters between two groups. Methrotra & Gilbert (2006) evaluate a variety of tests for an HIV vaccine trial, including separate tests of the proportion infected and of the post-

infection viremia as well as various weighted versions. For their setting, they conclude that separate tests have appeal. To deal with potential selection bias of the test in the infecteds, they suggest performing the sensitivity analyses using the approach of GBH and HHS.

In this paper, we introduce tests that have good power when two distributions differ in terms of the right tail of the distribution. With this procedure, the Ws are first sorted in each group and the group with the smallest number of zeros is determined. All zeros from this group are tossed out and the same proportion of zeros are chopped off the second group. A test statistic is created using the lump of data that remains. A permutation approach, where the entire data set is scrambled, provides a valid null distribution for the test statistic. These "chop-lump" tests are valid under the null hypothesis that the distribution of W is the same for the two groups, and have good power when the two distributions differ in the right tail. They can be viewed as a way to compare two truncated distributions, or an improved BOI test for lumpy skewed data. They can also be interpreted as a test in the principal stratum of the non-immune under certain assumptions. Interestingly, we prove that the chop-lump Wilcoxon test is asymptotically more powerful than the usual Wilcoxon when the infection rates are the same. This result is of interest beyond vaccine trials. The tests are evaluated by simulation for scenarios reflecting an HIV vaccine trial and a malaria vaccine trial. Finally, the methods are used to provide a re-analysis of the first phase III HIV vaccine trial (rgp120 Study Group 2004).

2 Tests in Vaccine Trials

To develop some notation, suppose that n patients are each randomized to vaccine (placebo) and that infections are recorded along with a measure of disease burden in the infecteds. Let Y denote the indicator of infection, $Z = C, V$ identify the group, and let X be a measure of disease burden e.g. viral load. We define W as 0 for an uninfected patient and X for an

infected patient, and have W_{Zi} ($W_{Z(i)}$) as the ith measurement (ordered measurement) in group Z. For notational ease, the treatment of unequal group sizes is deferred to Appendix A.

Exploiting the idea of Principle Stratification (Frangakis and Rubin 2002), Gilbert, Bosch, and Hudgens (2003) noted that for a vaccine trial, each subject enrolled in the trial must be one of four unknowable types, given in Table 1. For example, the subgroup of the "harmed" consists of those who would become infected if given the vaccine, but who would remain uninfected if given the placebo.

Using the observed data, we can straightforwardly test the null hypothesis that the distribution of W is the same in the two groups.

$$H_0^W : F_V(w) = F_C(w)$$

where $F_Z(w)$ is the cdf of W in group $Z = C, V$. Though unnecessary, note that the distribution of W can be decomposed in terms of the principle strata:

$$\begin{aligned} F_V(w) &= \delta_0(w)(\theta_{00} + \theta_{01}) + F_V(w|10)\theta_{10} + F_V(w|11)\theta_{11} \\ F_C(w) &= \delta_0(w)(\theta_{00} + \theta_{10}) + F_C(w|01)\theta_{01} + F_C(w|11)\theta_{11} \end{aligned}$$

where $F_Z(w|ij)$ is the distribution of W in the principle stratum $Y(v), Y(p) = i, j$ for $Z = V, C$ and $\delta_0(w)$ is the indicator that $w = 0$. Thus in the vaccine group, the 0s are a mixture of the immune and protected, while the vaccine Ws > 0 are a mixture of the harmed and doomed. A similar decomposition holds for the placebo group.

To test H_0^W, Chang, Guess, & Heyse (1994) suggested using the standardized difference in sample means of W:

$$Z_{BOI} = \frac{\sum_{i=1}^n W_{Ci}/n - \sum_{i=1}^n W_{Vi}/n}{\sqrt{2\overline{X}^2 \bar{p}(1-\bar{p})/n + \bar{p}(S_C^2/n + S_V^2/n)}}, \tag{1}$$

where $\bar{p}$ is the proportion of infections in the combined sample, $\overline{X}$ the overall mean of the Xs, and S_Z^2 the sample variance of X in group Z. On the null hypothesis, this has an asymptotic standard normal distribution. Z_{BOI} has the same numerator as the usual t-test, but a different variance estimate that exploits the semi-continuous nature of W. One can show that Z_{BOI} is asymptotically equivalent to the t-test under local alternatives.

While appealingly consistent with the intent to treat principle, the burden of illness test can have poor power if there are a substantial number of zeros. Intuitively this makes sense. One way to obtain a substantial number of zeros is to enroll many from the "immune" stratum. Such patients cannot receive a benefit from the vaccine and thus a trial with lots of zeros may have poorer power than a trial with few zeros.

If the vaccine does not prevent infections but improves X, a more powerful test can be fashioned by just using X from the infecteds. Under the strong null that the vaccine is inert, we have

$$H_0^S : \theta_{10} = \theta_{01} = 0, \quad F_V(x|11) = F_C(x|11),$$

where $F_Z(x|11)$ is the distribution of X in the doomed following randomization to Z. Note that this is stronger and different than H_0^W which permits $\theta_{10} = \theta_{01} > 0$. Under H_0^S, those who are infected must be from the doomed stratum and thus a valid test of H_0^S is obtained by comparing the distribution of X in the placebo infecteds to the distribution of X in the vaccine infecteds. For example, a t-test in the infecteds can be performed:

$$Z_{\text{inf}} = \frac{\sum_{i=1}^{m_C} X_{Pi}/m_C - \sum_{i=1}^{m_V} X_{Vi}/m_V}{\sqrt{S^2(1/m_C + 1/m_V)}},$$

where S^2 is the pooled sample variance of the Xs and m_Z is the number infected in group Z. While the n placebo (vaccine) recipients are ensured to be comparable in terms of seen and unseen baseline characteristics by randomization, the m_C (m_V) placebo (vaccine) infecteds are comparable only under the strong null. By comparable we mean that the infecteds come

from the same principal stratum; under H_0^S, necessarily the doomed. If H_0^S is not true, placebo infecteds are a mixture of the doomed and protected while the vaccine infecteds are a mixture of the doomed and harmed. Suppose the vaccine protected no one, caused infections with low viral loads, and had no impact on viral loads for the doomed. Then H_0^S is not true $\theta_{01} = 0$, $\theta_{10} > 0$, the observed vaccine viral loads would tend to be smaller than the observed placebo viral loads, and using Z_{inf} would tend lead to the catastrophically wrong conclusion that a pernicious vaccine is good. An illustration of this scenario is given in Figure 1.

Gilbert, Bosch & Hudgens (2003) addressed the problem of selection bias by testing vaccine efficacy in the principal stratum of the doomed. Within any principle stratum, the vaccine and control groups are comparable by virtue of randomization, so there is no issue about selection bias. They assume that $\theta_{10} = 0$, i.e. that the vaccine cannot harm anyone. Thus the control infecteds must be a mixture of the protected and doomed while the vaccine infecteds are purely the doomed. They propose culling out a proportion (based on a specified $\theta_{01}/(\theta_{11} + \theta_{01})$) of placebo infecteds who belong to the protected principle stratum. The remaining placebo people must be the doomed and their viral loads are compared to the viral loads in the vaccine infecteds (by assumption the doomed). There are different ways to do this culling and GBH propose a model for culling indexed by a single parameter, β, that reflects the degree of selection bias. If the test remains significant under different culling mechanisms, one might feel more comfortable concluding vaccine works in the doomed. Of course the most extreme culling has appeal, as this obviates the need to specify β.

Hudgens, Hoering & Self (2003) explicitly discuss the most extreme culling. Of particular interest is the scenario that penalizes the vaccine arm the most. It could be that the largest viral loads in the placebo group are all from the protected principal stratum. That is, the vaccine protects those who would have had the largest viral loads, if given placebo. If so,

then if we want to estimate the effect of vaccine in the estimated principal stratum of the doomed, we should toss out the largest viral loads among the placebo infecteds until we have m_V viral loads remaining. *Provided* $\hat{p}_V < \hat{p}_C$, their test statistic becomes

$$T_M = \frac{\sum_{i=1}^{m_V} X_{C(i)} - \sum_{i=1}^{m_V} X_{Vi}}{m_V},$$

for the difference in means approach, where the $X_{C(i)}$s are ordered from smallest to largest. If $\hat{p}_V \geq \hat{p}_C$, then (with the assumption that $\theta_{10} = 0$) the infecteds from each group are presumably from the doomed principal stratum. The fact that more infections occurred in the vaccine group is because by chance more doomed were randomized to vaccine than placebo. Thus the mean difference in viral load among the doomed is estimated as the mean difference in the infecteds:

$$T = \sum_{i=1}^{m_C} X_{Ci}/m_C - \sum_{i=1}^{m_V} X_{Vi}/m_V = \overline{X}_C - \overline{X}_V.$$

Note that if $p_C = p_V$, then about half the time T will be the mean viral load difference in the infecteds, reinforcing the importance of the stated assumption that $\theta_{10} = 0$ for this method to work. To obtain a null distribution they simulate data using a bootstrap procedure that simulates data from the NPMLEs of $F_V(w)$ and $F_C(w)$ assuming $\theta_{10} = 0$, the null hypothesis, and the most extreme selection model.

Another approach is to compare the infection rates between the two groups:

$$Z_{\text{prop}} = \frac{\hat{p}_C - \hat{p}_V}{\sqrt{2\overline{p}(1-\overline{p})/n}}$$

where $\overline{p} = (\hat{p}_C + \hat{p}_V)/2$. For a vaccine with little effect on acquisition, this would have poor power.

Lachenbruch (2001) proposed combining Z_{prop} with Z_{inf} by adding their squared values together. Under H_0^S, this has a chi-square distribution with 2 degrees of freedom. While this procedure is useful in certain settings, power will be compromised relative to a test

that focuses on the Xs if there is little effect on acquisition. Additionally, Lachenbruch points out that the power advantages of this procedure are greatest if the vaccine effects are contradictory e.g. reduces acquisition but increases viral load. Mehrotra & Gilbert (2006) introduce a variety of ways to combine Z_{prop} and Z_{inf}, including weighted tests. However, if there is expected to be little to no effect on acquisition, the optimal weighted test will assign 0 weight to Z_{prop} and we end up with a comparison of the infecteds. They suggest conducting separate tests Z_{prop} and Z_{inf} with an investigation of selection bias for Z_{inf} using the methods of GBH.

3 Chop-Lump Tests

Our goal is to develop a more powerful BOI-like test that will not result in the wrong conclusion with a harmful vaccine, and to have good power for trials where the vaccine should have little effect on acquisition. For a vaccine with no acquisition, the proportion of zeros should not be informative about the vaccine and it makes sense to focus on the tail of the distribution where the numbers reside. Towards this end, we propose a "chop-lump" procedure. Under this approach, the data are sorted separately and we throw out an equal proportion of zeros in each group so that one group has no zeros. This even-handed approach is illustrated in Table 2. Different test statistics can be constructed using the data to the right of the chopping point. A null distribution can be obtained by permutation where the entire data set is scrambled, a new chopping point determined, and the test statistic re-constructed.

Similar to the BOI, test, we can calculate a standardized mean difference in the remaining Burden-of-Illness scores by using the Ws to the right of the chopping point.

$$\frac{\sum_{i=(n-m)+1}^{n} W_{V(i)}/m - \sum_{i=(n-m)+1}^{n} W_{C(i)}/m}{\sqrt{2S_m^2/m}}$$

where S_m^2 is the pooled sample variance based on the m largest Ws in each group, and $m =$

$\max(m_C, m_V)$. Note that though a permutation null distribution is used, the denominator is not superfluous as m and thus S_m^2 will change in the permuted datasets. We call this the chop-lump t-test (CLT).

We also propose a standardized rank test which we call the chop-lump Wilcoxon (CLW)

$$\frac{\sum_{i=n-m}^{n} \text{rank}(W_{V(i)}) - m(2m+1)/2}{m^2(2m+1)/12 - m(L^3 - L)/\{24(m-1)\}}$$

where L is the number of leftover zeros and $\text{rank}(W_{V(i)})$ is the (mid)-rank of $W_{V(i)}$ in the lump of data to the right of the chopping point

$$W_{V(n-m+1)}, W_{V(n-m+2)}, \ldots, W_{V(n)}, W_{C(n-m+1)}, W_{C(n-m+2)}, \ldots, W_{C(n)}.$$

Appendix A gives details on how to efficiently obtain the exact permutation distribution of both CLT and CLW when the number of $W_i > 0$ is small and gives an approximation otherwise. The approximation relies on a (hypergeometric) weighted sum of Gaussian cumulative distribution functions. A small simulation is performed to demonstrate the accuracy of the approximation. The **R** package choplump is available at **http://cran.r-project.org/** to perform both chop-lump tests.

Getting back to Figure 1, note that for this pernicious vaccine, the Ws to the right of the chopping point are given by the light and dark shaded areas. Note that the median (mean) burden of illness to the right of the chopping point is equal (smaller) in the placebo group compared to the vaccine group. Thus the chop-lump tests should not tend to conclude a harmful vaccine is good. The chop-lump tests are also more appealing when selection bias goes the other way. If we switch the placebo and vaccine labels in Figure 1, then the light shaded group corresponds to the protected principle stratum and the vaccine is protecting those who would have had infections with low viremia, if given placebo. A test in the infecteds tends to conclude that a good vaccine is bad while the chop-lump tests appropriately favor the vaccine group.

Now imagine placing the lightly shaded small hump in Figure 1 to the far right so the vaccine is causing infections with high viremia. Here both the chop-lump and test in the infecteds appropriately have higher means for the vaccine group. If we then imagine switching the vaccine and placebo labels, so the vaccine is preventing infections with high viremia, the test in the infecteds would tend to draw the correct conclusion as would the chop-lump test. Thus for all these extreme scenarios, the chop-lump tests tend to draw the correct conclusions. This is not true for the test in the infecteds, which can tend to conclude a harmful vaccine is good or a good vaccine is harmful.

Under certain assumptions, the chop-lump tests can be viewed as a test of vaccine effectiveness, as measured by W, in the principle stratum of the non-immune i.e. the union of the doomed, harmed, and protected principle strata. Note that without making any assumptions, we can write the distribution of W in the nonimmune groups as

$$F_V^{NI}(w) = F_V(w|1)\frac{\theta_{10}+\theta_{11}}{\theta_{10}+\theta_{01}+\theta_{11}} + \delta_0(w)\frac{\theta_{01}}{\theta_{10}+\theta_{01}+\theta_{11}} \quad (2)$$

$$F_C^{NI}(w) = F_C(w|1)\frac{\theta_{01}+\theta_{11}}{\theta_{10}+\theta_{01}+\theta_{11}} + \delta_0(w)\frac{\theta_{10}}{\theta_{10}+\theta_{01}+\theta_{11}}. \quad (3)$$

where $F_Z(w|1)$ is the distribution of W for the infecteds in group z, and is readily estimable. Note that if θ_{01}, θ_{10} and θ_{11} were estimable, we could estimate the mixing proportions and thus estimate $F_Z^{NI}(w), Z = C, V$.

While these mixing proportions are not estimable, note that without making any assumptions we have

$$p_C = \theta_{01} + \theta_{11}$$

$$p_V = \theta_{10} + \theta_{11}$$

thus we can deduce that θ_{11} must lie within the interval $[0, \min(p_C, p_V)]$. If θ_{11} were known, we could estimate θ_{01}, θ_{10}, using the above equations and thus estimate $F_Z^{NI}(w)$ and construct

a test. With θ_{11} unknown, we could construct a family of tests for all $\theta_{11} \in [0, \min(\hat{p}_V, \hat{p}_C)]$. Note that if $\theta_{11} = 0$ then a test in the estimated nonimmune stratum would be based on the Ws remaining after we throw out the smallest $n[1 - (\hat{p}_C + \hat{p}_V)]$ Ws from each group. On the other hand if $\theta_{11} = \min(\hat{p}_C, \hat{p}_V)$, then a test in the estimated nonimmune stratum corresponds to throwing out the smallest $n[1 - \max(\hat{p}_C, \hat{p}_V)]$ Ws from each group—exactly the chop-lump procedure of Table 2.

The chop-lump tests can be viewed as tests in the estimated principal stratum of the nonimmune under an awkward assumption–loosely that we assume that either the protected or the harmed group don't exist, but we don't know which. Note that the selection bias approaches of HHS and GBH made the assumption that the harmed group didn't exist. So in some sense, the chop-lump approach is an even-handed generalization of HHS and GBH. More formally, we can obtain the mles of $\theta_{01}, \theta_{10}, \theta_{11}$ under a restricted parameter space $\mathcal{R} = \{(\theta_{01}, \theta_{10}, \theta_{11}) : [(\theta_{01}, \theta_{10}) \propto e_1 \bigcap \theta_{11} \in [0,1], \theta_{01} + \theta_{10} + \theta_{11} \leq 1] \quad \text{OR} \quad [(\theta_{01}, \theta_{10}) \propto e_2 \bigcap \theta_{11} \in [0,1], \theta_{01} + \theta_{10} + \theta_{11} \leq 1]\}$, where $e_1 = (1,0)$ and $e_2 = (0,1)$. Note that $\mathcal{R}$ is a subset of two planes in R^3. We can then replace $F_Z(w|1)$ with its NPMLE and form $\hat{F}_Z^{NI}(w)$ by weighting the $\hat{F}_Z(w|1)$s and $\delta_0(w)$s by the restricted mles of $\boldsymbol{\theta}$ according to the formulas (2) and (3). Importantly, the validity of the chop-lump procedure does not depend on a principal stratum interpretation.

4 Asymptotic Representation of Power

In Appendix B , we derive asymptotic expressions for the power of the different tests under local alternatives. We assume that the Xs follow a distribution with $\Delta = E[X_C] - E[X_V]$,var$[X_C] =$ var$[X_V] = \sigma^2$, $P(Y = 1|Z = C) = p_C$ and $P(Y = 1|Z = V) = p_V$, and that n patients are randomized per group. Denote $p = (p_C + p_V)/2$. The use of local alternatives means that Δ, and $p_C - p_V$ vary with n so that the expected value of the standardized

test statistic converges to a constant.

It is straightforward to show that the asymptotic power function for the BOI t-test, under local alternatives, is

$$\Phi(\mathrm{E}(Z) - z_{\alpha/2}), \tag{4}$$

where

$$E[Z] = \sqrt{n}\frac{p_V\Delta + E[X_C](p_C - p_V)}{\sqrt{2\{p_V/(p_C + p_V)\Delta + E[X_C]\}^2 p(1-p) + 2p\sigma^2}}$$

is the expected value of the BOI t-test. Note that if $p_C = p_V$, increasing $E[X_C]$ decreases power. Thus the "gap" between 0 and the mean Burden of Illness has an important effect on power.

While, as we will see, the power for the chop-lump t-test can be substantially greater than the usual t for realistic situations, one can show that the two tests are asymptotically equivalent (Appendix B). The speed of convergence of this approximation for the chop-lump t-test can be quite slow if $E[X_C]$ is large, or if p is small with the chop-lump t having better power than the usual t-test.

While the powers of the Chop-Lump and BOI t-tests unappealingly depend on $E[X_C]$, the powers of the rank tests do not. The asymptotic power, under local alternatives, for the Wilcoxon test corrected for ties is

$$\Phi\left[\frac{\sqrt{\theta}\,p\mathrm{E}(Z_{\mathrm{inf}}) + \sqrt{3(1-p)}\,\mathrm{E}(Z_{\mathrm{prop}})}{\sqrt{(1-p)^2 + (1-p) + 1}} - z_{\alpha/2}\right], \tag{5}$$

where

$$\begin{aligned} E[Z_{\mathrm{inf}}] &= \frac{\Delta}{\sqrt{2\sigma^2/(np)}}, \\ E[Z_{\mathrm{prop}}] &= \frac{p_V - p_C}{\sqrt{2p(1-p)/n}}, \end{aligned}$$

are the expected values of the t-test on the infecteds, and the tests of proportion infecteds, respectively and $\theta = 3/\pi$ is the asymptotic relative efficiency (ARE) of the Wilcoxon test

compared to a t-test for underlying Gaussian data. If X followed a different distribution, θ would correspond to the associated ARE.

The asymptotic power expression, under local alternatives, for the Chop-Lump Wilcoxon test is

$$\Phi\left[\frac{\sqrt{\theta}\,\mathrm{E}(Z_{\mathrm{inf}})+\sqrt{3(1-p)}\,\mathrm{E}(Z_{\mathrm{prop}})}{\sqrt{1+3(1-p)}}-z_{\alpha/2}\right] \tag{6}$$

We first compare the terms inside the power expressions for chop-lump rank and the Wilcoxon test. If the vaccine has no effect on acquisition, we have $p_C = p_V = p$ and $E[Z_{\mathrm{prop}}] = 0$. It is easy to show in this case that the chop-lump rank test always has equal or greater power than the Wilcoxon rank test. On the other hand, if the vaccine has no effect on viremia in the infecteds, then the Wilcoxon rank test is never worse. But in this setting, we should really be doing just a test of proportions.

More generally, we can equate the two terms and calculate the point where power for the two Wilcoxon tests is the same:

$$E[Z_{\mathrm{inf}}]/E[Z_{\mathrm{prop}}] = \sqrt{\frac{3q}{\theta}}\frac{(\sqrt{1+3q}-\sqrt{q^2+q+1})}{(\sqrt{q^2+q+1}-(1-q)\sqrt{1+3q})}$$

where $q = 1-p$.

Figure 2 plots the curve of indifference between the chop-lump Wilcoxon and the usual Wilcoxon tests. As the proportion infected gets smaller, the chop-lump Wilcoxon test has better power for smaller ratios $E[Z_{\mathrm{inf}}]/E[Z_{\mathrm{prop}}]$. As $q \to 1$ the ratio approaches $(2-\sqrt{3})/\sqrt{\theta}$.

We can also compare the chop-lump Wilcoxon to the BOI-t test when $p_C = p_V = p$. In this setting one can deduce that the chop-lump rank test is preferred to the BOI-t test provided

$$\sqrt{\theta}\sqrt{2\sigma^2 p + 2(\Delta/2 + E[X_C])^2 p(1-p)} > \sqrt{2\sigma^2 p + 6\sigma^2 p(1-p)}$$

For fixed $E[X_C]$, we have $\Delta \to 0$, as $n \to \infty$ for local alternatives. Thus for a distribution

with $\theta = 1$ the chop-lump rank test is asymptotically preferred to the BOI t-test provided

$$\frac{E[X_C]}{\sigma} > \sqrt{3}.$$

So if the gap is about 1.7 or more standard deviations, the chop-lump rank test is asymptotically preferred. Interestingly, this does not depend on p—the proportion infected.

In small samples, of course, the asymptotic power expressions may not be accurate and simulation may be required to obtain accurate probabilities.

5 HIV and Malaria Vaccine Trials

5.1 HIV Vaccine Trial

To evaluate the performance of these tests in realistic conditions, we consider vaccine trials of HIV and malaria. Current HIV vaccine candidates are expected to work by inducing cell mediated immunity, resulting in T-cells that are primed to kill HIV-infected cells. But the vaccine may have little to no effect on acquisition of infection. HIV vaccine trials need to enroll large numbers of patients, as the infection rate is relatively rare. We first consider a scenario intended to mimic PAVE 100, an HIV vaccine trial that will evaluate a NIAID vaccine. PAVE 100 will be a randomized, placebo controlled trial of an HIV vaccine in 8500 individuals, designed to accrue 180 infections in roughly equally from three regions of the world.

In the placebo group, the number of infections were generated according to a binomial distribution with $N = 4250$ and $p_C = 90/4250$. The vaccine group used $p_V = (90/4250)(1 - \mathrm{VE}_S)$, where the vaccine efficacy on susceptibility or VE_S ranged from 0 to .2. For infected participants, log10 viral loads were generated according to a normal distribution with variance $.75^2$ and mean 4.5 or $4.5 - \Delta$ for patients in the placebo or vaccine group based on data from the Multi-Center AIDS Cohort Study (Lyles et al 2000). We also evaluated a pernicious vaccine which does nothing other than cause some infections with low

levels of viremia. Here we have $\theta_{10} > 0$ with $p_C = \theta_{11} = 90/4250$. Thus $p_V = 90/4250 + \theta_{10}$. Within the doomed (harmed) stratum X was normal with mean 4.5 (2.5) and variance $.75^2$.

We evaluated the BOI test (1) and the rank version of the BOI (the usual Wilcoxon test) on all the data. We also evaluated the chop-lump t and Wilcoxon tests as well as a t-test in the infecteds and a rank version of the HHS test in the doomed under selection bias that maximally penalizes the vaccine. For each scenario, 1,000 clinical trials were simulated. Permutation null distributions were simulated for the rank test in the doomed and the chop-lump tests using 299 permutations. For the remaining tests, a standard normal null reference distribution was used. The Wilcoxon test was standardized using a variance that took into account the number of tied observations (see Lehmann 1975). All tests are one-sided with $\alpha = .025$. Table 3 presents the results.

The first row indicates that all tests control the type I error rate. Rows 2-4 are for a vaccine with a modest effect on viremia but with VE_Ss of, respectively, 0%, 10%, and 20%. We see that with a VE_S of 0%, the test in the infecteds has the best power with the rank test in the doomed somewhat worse. The chop-lump Wilcoxon test does the best of the remaining tests but is substantially worse than the test in the infecteds and the test in the doomed. The power of the test in the infecteds remains similar for increasing VE_Ss, but decreases for the rank test in the doomed. This is because the benefit of the vaccine on acquisition is not reflected in the doomed. Further, as VE_S increases, more of the largest viral loads in the placebo group are thrown out. The chop-lump and BOI tests both increase in power with increasing VE_Ss. The effect of VE_S is pronounced for all tests, except the t-test in the infecteds. Rows 5-7 are similar but with a more pronounced vaccine effect of 1.0 logs, and the relative conclusions are the same as for rows 2-4. However, we do see that for this alternative, the CLW test has power $> 90\%$ for all 3 VE_Ss. The 8th row is for the hypothetical case when $E[X_C]$ is 1.5 instead of 4.5 so the gap is about 2 standard deviations

instead of about 6. Here the powers for the chop-lump t-tests and the BOI t-test are similar, reinforcing the importance of the "gap" ($E[X_C]$) on power. Finally, the last two rows show a scenario with a pernicious vaccine that does nothing but cause either 10% or 20% more infections in the vaccine group with a low mean viremia of 2.5 logs. Tests in the infecteds have 30% to 73% power to draw a catastrophically wrong conclusion, that the vaccine is beneficial on mean viral load when it is actually causing infections. The test in the doomed is not designed for this scenario with $\theta_{10} > 0$ and also tends to conclude a harmful vaccine is beneficial. For PAVE 100, the chop-lump Wilcoxon will be an important test to evaluate.

5.2 Malaria Vaccine Trial

Malaria is quite different from HIV. In endemic areas, individuals may be repeatedly infected by mosquitoes who transfer the malaria parasites to the human during a blood meal. These parasites then travel to the liver where they mature, and later infect red blood cells. Following further development in the red blood cells, they ultimately are taken in by some mosquito during a blood meal weeks to months after the initial blood meal. Current vaccine candidates at NIAID are designed to induce antibodies to antigens that are exposed on the parasite just prior to docking and infection of red blood cells. Thus these vaccines are not expect to have an impact on the acquisition of infection or on the liver stage of the parasite's life cycle (Girard et al 2007). Since parasites proliferate and then burst out of infected red blood cells in huge numbers, if red blood cell infection is prevented, there should be less parasitemia or parasites in the bloodstream.

It is likely that volunteers from endemic areas will have been infected multiple times prior to trial enrollment and the proportion with detectable parasitemia (parasites in the blood) can be 90% or greater. Hopefully the vaccine will reduce the parasitemia. A typical phase II malaria trial with a parasitemia endpoint might enroll about 150 volunteers per group. Data

from Smith, Schellenberg, & Hayes (1994) of a study of 426 children from Tanzania suggest that the distribution of log10 parasitemia has a mean of about 3.5 logs and a variance of approximately $(1/3)^2$ which we assume follows a normal distribution for these simulations. Approximately 90% of the children were infected. Results from simulating such trials under different scenarios are provided in Table 4.

As in Table 4, we see that all tests control the type I error rate. The test in the infecteds has about 96% power for the scenarios in rows 2-4, while the power of the test in the doomed decreases here as the VE_S increases. Among the other tests, the Wilcoxon tests have similar power and perform better than the t-tests. While the asymptotic power expressions suggest the chop-lump Wilcoxon should have larger power than the ordinary Wilcoxon, the magnitude of the power differential is quite modest for this scenario. With $E[X_C] = 1.5$, the power of the BOI and chop lump t-tests increase and are more similar compared to $E[X_C] = 3.5$, As before, the test in the infecteds can lead to the conclusion that a harmful vaccine is beneficial, as can the test in the doomed. Thus for this scenario, either the usual or chop-lump Wilcoxon are appealing: good to best power coupled with desirable behavior for a vaccine that causes low parasitemia infections. The t-tests have relative poor power here—note that the gap between the zeros and numbers is roughly 10 standard deviations. Nonetheless, the chop-lump t-test is more powerful than the usual t-test here.

6 Example: VAX004

VAX004 was the first phase III trial of an HIV vaccine. A total of 5403 volunteers at high risk of acquiring HIV via sexual transmission were randomized in a 2:1 ratio to vaccine and placebo (rgp 120 HIV Vaccine Study Group 2004). Volunteers received 7 injections spaced 6 months apart. It was hoped that the vaccine would induce antibodies that would prevent infection, and acquisition of HIV infection was the primary endpoint. The trial demonstrated

that the vaccine was ineffective. The overall placebo and vaccine infections rates were 6.8% and 6.2% respectively, with an estimated percent reduction of the infection rate of 6% with a 95% CI of -17% to 24%.

Following detection of infection, volunteers were followed to evaluate the progression of HIV disease. While this vaccine was not designed to invoke cell mediated immunity—where the immune system kills infected cells—serial measurements of HIV viral load, CD4 counts, and ART initiation were obtained to evaluate post-infection outcomes. Here we use the results of VAX004 to highlight application of the chop-lump procedures on viral load.

The first two panels of Figure 3 provide the distribution of W, which is the average of the first two (log10) viral loads following detection of infection for the infecteds, and 0 for the uninfecteds. We see that there is a substantial proportion of zeros in each group and that there is a large gap between the 0s and $E[X_C] = 4.2$, as $\text{std}[X_C] = .84$. Approximately 5 standard deviations separate the mean placebo viral load from 0. Thus the chop-lump Wilcoxon test proposed in this paper would seem especially suitable for analyzing these data.

The bottom two panels of Figure 2 show the right tails of the distribution of W after chopping, along with the median viral load. The two medians are virtually the same, 4.15 for the placebo and 4.19 for the vaccine group. The chop-lump permutation p-value for the Wilcoxon (t) test is .69 (.56), indicating no effect of vaccine on the 'burden of illness" in the right tail of the distribution of W.

7 Discussion

For a vaccine that is intended to reduce disease burden but have little to no effect on acquisition, deciding on an analysis strategy is complicated. While the BOI tests are conceptually appealing for this purpose as they use all the data, they can have poor power, particularly

if the gap between 0 and $E[X_C]$ is large. Testing equality of the distribution of X in the infecteds can make sense for non-licensure trials provided the vaccine has a remote likelihood of affecting acquisition as they can have substantially more power than the chop-lump tests and BOI tests. However, the test in the infecteds can lead to a catastrophically wrong conclusion under a pernicious vaccine, as can tests in the principal stratum of the doomed. Especially for licensure trials, use of a test that avoids these problems is appealing.

While our focus has been on vaccine trials, the results of this paper apply more generally to comparing two groups with semi-continuous or mixed discrete and continuous outcomes, where the discrete mass of data is all at one tail. Of particular interest are the conditions, graphed in Figure 1, where the chop-lump Wilcoxon has better asymptotic power than the usual Wilcoxon.

ACKNOWLEDGMENTS

We thank Peter Gilbert and Bryan Shepherd for helpful comments on this manuscript. We thank Marc Gurwith and the Vaxgen corporation for kind use of the VAX004 trial data.

REFERENCES

Alexander LN, Seward JF, Santibanez TA, Pallansch MA, Kew OM, Prevots DR, Strebel PM, Cono J, Wharton M, Orenstein WA, & Sutter RW (2004) "Vaccine Policy Changes and Epidemiology of Poliomyelitis in the United States." *Journal of the American Medical Association.* 292 (14) 1696-1701.

Burke DS (1992) "Human HIV vaccine trials: does antibody dependent enhancement post a genuine risk?" *Perspectives in Biology and Medicine* 35 511-530.

Chang MN, Guess HA, Heyse JF (1994) "Reduction in burden of illness: a new efficacy measure for prevention trials," *Statistics in Medicine* 13 8 1807-1814.

Frangakis CE, Rubin DB, (2002) "Principal stratification in causal inference," *Biometrics* 58 (1): 21-29.

Gilbert PB, Bosch RJ, Hudgens MG (2003) "Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials," *Biometrics* 59 (3): 531-541.

Gilbert PB, Ackers ML, Berman PW, Francis DP, Popovic V, Hu DJ, Heyward WL, Sinangil F, Shepherd BE, & Gurwith M (2005) "HIV-1 virologic and immunologic progression and initiation of antiretroviral therapy among HIV-1 - Infected subjects in a trial of the efficacy of recombinant glycoprotein 120 vaccine," *Journal of Infectious Diseases* 192 (6): 974-983.

Girard MP, Reed ZH, Friede M, & Kieny MP (2007) "A review of human vaccine research and development: Malaria," *Vaccine* 25(9): 1567-1580.

Hudgens MG, Hoering A, Self SG (2003) "On the analysis of viral load endpoints in HIV vaccine trials," *Statistics in Medicine* 22 (14): 2281-2298.

Lachenbruch PA (2001) "Comparisons of two-part models with competitors," *Statistics in Medicine* 20 8 1215-1234.

Lehmann E (1975) "Nonparametrics: Statistical Methods Based on Ranks," Holden Day: Oakland, CA.

Lyles RH, Munoz A, Yamashita TE, Bazmi H, Detels R, (2000) "Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study" *Journal of Infectious Diseases* 181(3) 872-880.

Mehrotra DV, Li XM, Gilbert PB (2006) "A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial," *Biometrics* 62 (3): 893-900.

rgp 120 HIV Vaccine Study Group (2004) "Placebo-controlled phase 3 of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection," *Journal of Infectious Diseases* 191: 654-665.

Shepherd BE, Gilbert PB, Jemiai Y, Rotnitzky A (2006) "Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials" *Biometrics* 62 (2): 332-342.

Smith T, Schellenberg JA, Hayes R,(1994) "Attributable fraction estimates and case definitions for malaria in endemic," *Statistics in Medicine*, 13 22 2345-2358.

Tu W , Zhou, X-H (1999) "A Wald Tests comparing medical costs based on log-normal distributions with zero valued costs," *Statistics in Medicine*, 18: 2749-2761.

Table 1

Principal Strata for a vaccine trial. We can imagine that each volunteer must be one of four types, listed under the Stratum column. Here $Y(V), Y(C)$ *denote the infection indicators that would be observed if a person is randomized to vaccine, control, respectively. The random variables* $W(Z)$ *and* $X(Z)$ *are defined analogously, with* $X(Z)$ *a burden of illness measure, e.g. viral load, that is only observable in the infecteds.*

Y(V)	Y(C)	Stratum	W(V)	W(C)	Probability
0	0	immune	0	0	θ_{00}
1	0	harmed	X(V)	0	θ_{10}
0	1	protected	0	X(C)	θ_{01}
1	1	doomed	X(V)	X(C)	θ_{11}

Table 2

Data layout for chop-lump tests. Based on the sorted Ws, the two data vectors are chopped as soon as the zeroes stop. The lump of data to the right of the chopping point is used to compare the two groups. A valid permutation distribution obtains by repeatedly scrambling the vaccine/placebo labels on all the data, recalculating the chopping point, tossing out data to the left and recalculating the test using the data to the right.

														C							
Vaccine Ws :	0	0	0	0	0	0	0	0	0	0	0	0	0	**H**	0	0	1	2	4	8	9
Placebo Ws :	0	0	0	0	0	0	0	0	0	0	0	0	0	**O**	1	3	5	6	8	9	9
														P							

Table 3

Power for different tests of viremia and infection for a phase IIB HIV vaccine trial. A total of 8500 volunteers are randomized and a control infection rate of about 2.1% is anticipated.

			All Data		Chop-lump		t-test in	Rank test in
VE_S	$E[X_C]$	Δ	t-test	Wilcoxon	t-test	Wilcoxon	Infecteds	Doomed
0	4.5	0.0	0.026	0.028	0.024	0.023	0.025	0.027
0	4.5	0.4	0.087	0.026	0.175	0.390	0.890	0.814
10	4.5	0.4	0.259	0.077	0.371	0.629	0.868	0.634
20	4.5	0.4	0.515	0.221	0.619	0.809	0.869	0.409
0	4.5	1.0	0.373	0.025	0.601	0.938	1.000	1.000
10	4.5	1.0	0.644	0.082	0.850	0.990	1.000	0.999
20	4.5	1.0	0.852	0.233	0.954	0.998	1.000	0.997
0	1.5	0.4	0.415	0.060	0.403	0.388	0.890	0.812
-10	4.5	*	0.005	0.002	0.007	0.018	0.304	0.187
-20	4.5	**	0.002	0.000	0.007	0.012	0.728	0.512

* (**) - Vaccine causes a 10% (20%) increase in infections w/ mean viremia 2.5

Table 4

Power for different tests of parasitemia and infection for a phase II Malaria vaccine trial. A total of 300 volunteers are planned with an anticipated control infection rate of about 90%.

			All Data		Chop-lump		t-test in	Rank test in
VE_S	$E[X_C]$	Δ	t-test	Wilcoxon	t-test	Wilcoxon	Infecteds	Doomed
0	3.5	.00	0.025	0.027	0.021	0.025	0.023	0.018
0	3.5	.15	0.189	0.864	0.356	0.871	0.959	0.923
6	3.5	.15	0.619	0.953	0.764	0.953	0.953	0.786
11	3.5	.15	0.900	0.993	0.946	0.990	0.959	0.624
0	1.5	.15	0.575	0.864	0.626	0.871	0.959	0.923
*	3.5	**	0.000	0.019	0.000	0.015	0.589	0.294

** - Vaccine causes additional infections w/ mean parasitemia 2.5. All vaccinees are infected.

Figure 1: Distribution of W from a trial with a harmful vaccine whose only effect is to cause some infections with low viremia. Thus there are three principal strata: immune-clear, harmed-light shading, doomed-heavy shading. The mean viral load in the infecteds is lower in the vaccine group, which is comprised of the doomed & harmed, than in the placebo group who are only the doomed. An analysis using just the infecteds would tend to conclude that this pernicious vaccine is beneficial.

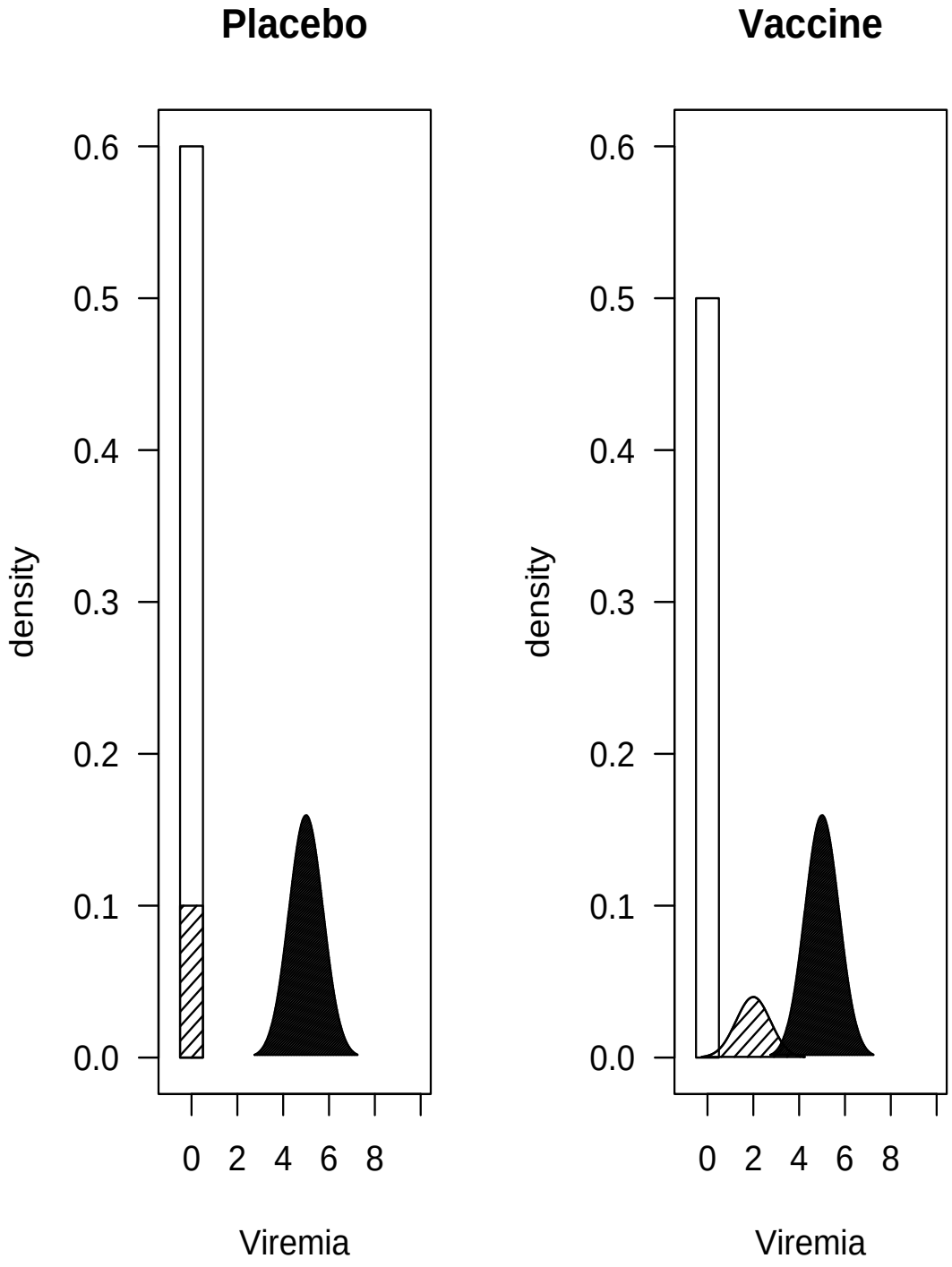


Figure 2: Boundary of indifference between the chop-lump and usual Wilcoxon tests as a function of the proportion uninfected and the ratio $E[Z_{\text{inf}}]/E[Z_{\text{prop}}]$. Note that if $p_V \to p_C$, and $E[Z_{inf}]$ is fixed, this ratio $\to \infty$ and the chop-lump Wilcoxon is always preferred.

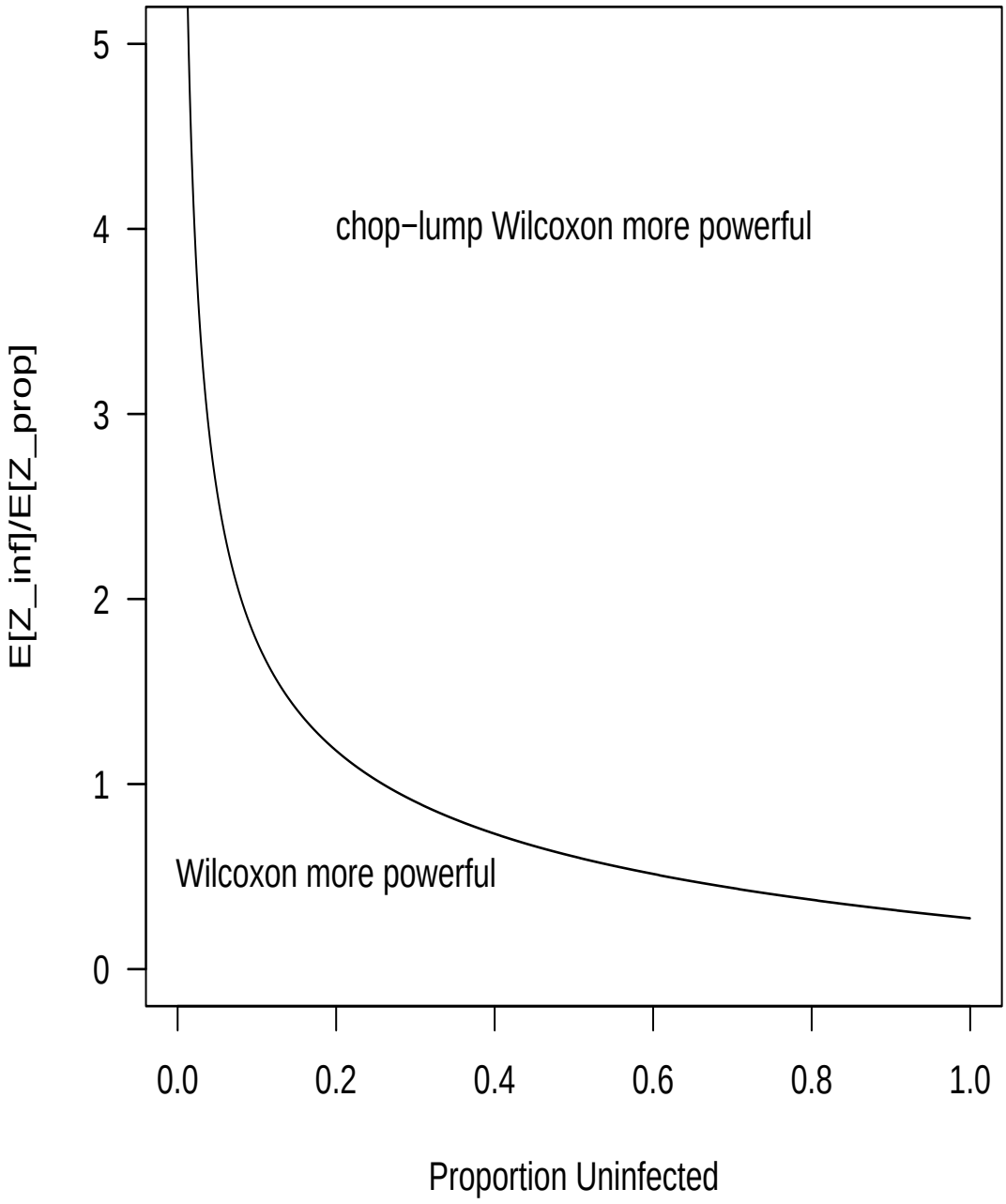
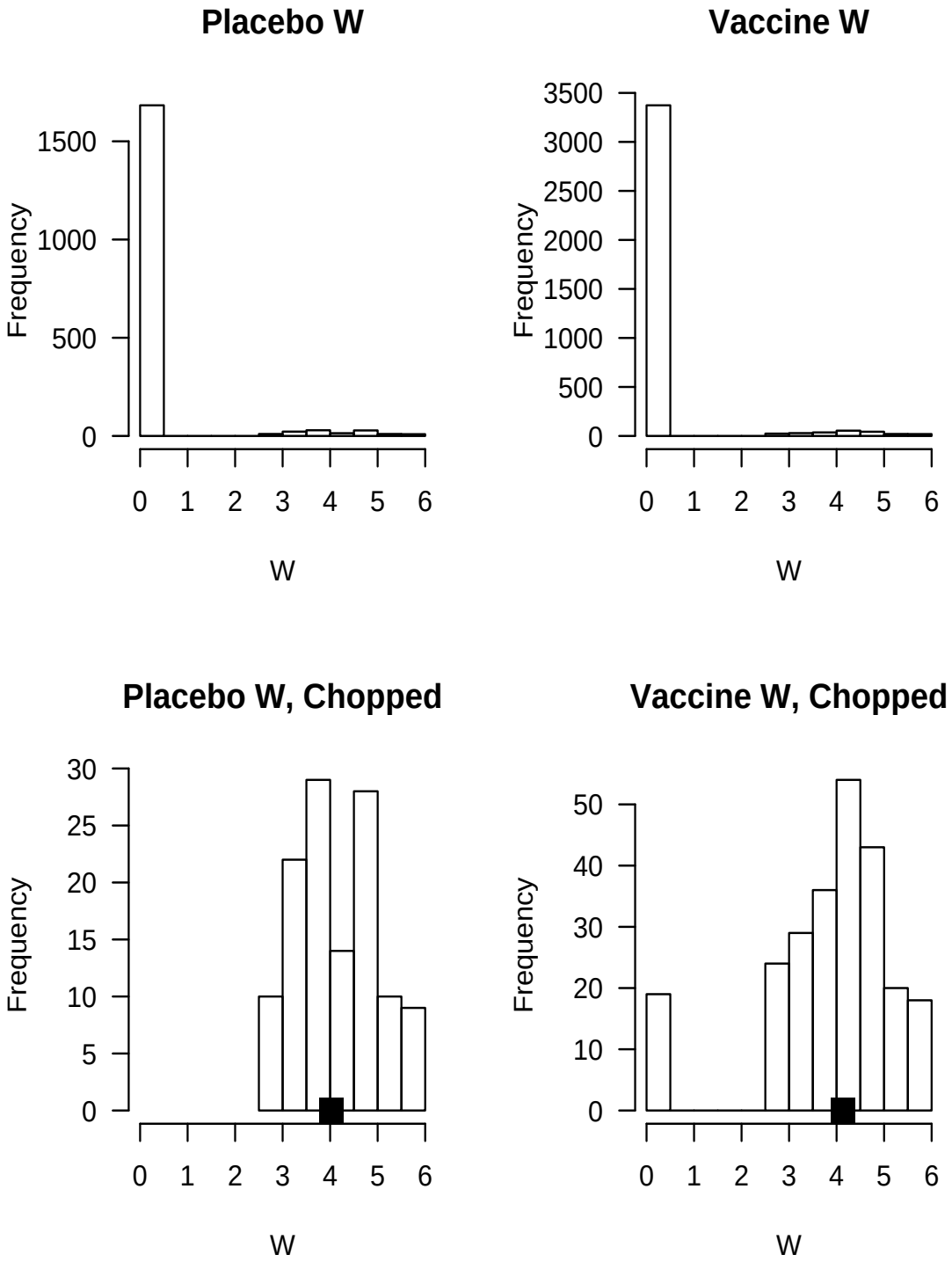


Figure 3: Histograms of W ($W = 0$ for uninfecteds and $W =$ viral load for infecteds). The upper panels provide the distribution based on all randomized participants, while the lower panels provide the distribution of W to the right of the chopping point.



Appendix A: Calculating p-values for chop-lump tests

General Chop-Lump Test

Suppose there are n_C and n_V subjects and m_C and m_V infections in the control and vaccine groups, respectively. Here we allow $n_C \neq n_V$, which causes some notational complexity, although the chopping function simply removes zeros in approximately the same proportion within each group such that one group has no zeros.

Here are the details. Let $k_i = n_i - m_i$, $N = n_C + n_V$, $M = m_C + m_V$, and $K = k_C + k_V$. Let the data be represented by two vectors, $\mathbf{W} = [W_1, W_2, \ldots, W_N]$ and $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, where Z is the indicator of vaccine. We order the indices $i = 1, \ldots, N$ by first ordering W_i and then by ordering Z_i within tied W_i values, so that $Z_1, \ldots, Z_{k_C}$ are zeroes and $Z_{k_C+1}, \ldots, Z_K$ are ones. Thus raw data $(0, 1, 0, 3, 0, 8)$ for group 0 and $(0, 5, 6, 4)$ for group 1 become $\mathbf{W} = (0, 0, 0, 0, 1, 3, 4, 5, 6, 8)$, $\mathbf{Z} = (0, 0, 0, 1, 0, 0, 1, 1, 1, 0)$. Here $(n_C, k_C, m_C) = (6, 3, 3)$ and $(n_V, k_V, m_V) = (4, 1, 3)$

Let $\mathbf{W}_a$ and $\mathbf{Z}_a$ be the last a values of $\mathbf{W}$ and $\mathbf{Z}$, respectively. Let $\mathbf{0}_a$ and $\mathbf{1}_a$ be vectors of zero or one of length a, where $a = 0$ denotes no vector (e.g., $[\mathbf{0}_3, \mathbf{1}_0]$ is a 3×1 vector of 0's). Let $C(\mathbf{W}, \mathbf{Z})$ be the chopping function which creates the "chopped" data set; specifically,

$$C(\mathbf{W}, \mathbf{Z}) = (\mathbf{W}_{M+a+b}, [\mathbf{0}_a, \mathbf{1}_b, \mathbf{Z}_M]) \,,$$

where

$$\text{if } \tfrac{m_C}{n_C} \geq \tfrac{m_V}{n_V} \qquad \text{then } a = 0 \text{ and } b = k_V - \lfloor \tfrac{n_V k_C}{n_C} \rfloor$$

and

$$\text{if } \tfrac{m_C}{n_C} < \tfrac{m_V}{n_V} \qquad \text{then } a = k_C - \lfloor \tfrac{n_C k_V}{n_V} \rfloor \text{ and } b = 0$$

and $\lfloor x \rfloor$ is the largest integer less than or equal to x. a and b are the numbers of remaining

zeroes in the control and vaccine groups, respectively, after chopping. Thus for our example dataset, $m_C/n_C = 3/6 < 3/4 = m_V/n_V$. Thus $a = 3 - \lfloor \frac{6\times 1}{4} \rfloor = 2$, $b = 0$; 2 of the 3 zeroes remain in the control arm and 0 of the 1 zero in the vaccine arm remain after chopping. We thus obtain $C(\mathbf{W}, \mathbf{Z}) = ([0, 0, 1, 3, 4, 5, 6, 8], [0, 0, 0, 0, 1, 1, 1, 0])$.

In the usual permutation test, we define a test statistic T which is a function of $\mathbf{W}$ and $\mathbf{Z}$. Let T_0 be the test statistic evaluated at the original data, and T_j be the test statistic evaluated at the jth permutation of the values of $\mathbf{Z}$. If lower values of the test statistic are more extreme, then a one-sided p-value is

$$p = \frac{\sum_{j=1}^{N!} I\left\{T_j \leq T_0\right\}}{N!} \tag{7}$$

where $I(a) = 1$ if a is true and 0 otherwise. A chop-lump test is simply a permutation test where the test statistic is of the form, $T_{CL}(\mathbf{W}, \mathbf{Z}) = T\left\{C\left(\mathbf{W}, \mathbf{Z}\right)\right\}$.

Exact p-values

In this section, we describe exact computation for any two-sample permutation test. There are computationally better ways to calculate the p-value than equation 7. First, we need not enumerate all $N!$ permutations of $\mathbf{Z}$, since there are only $\left(\begin{array}{c} N \\ n_V \end{array} \right)$ unique permutations of $\mathbf{Z}$, and each has exactly $n_C! n_V!$ permutations which correspond to the same permuted $\mathbf{Z}$. We can obtain similar computational savings by partitioning the $\left(\begin{array}{c} N \\ n_V \end{array} \right)$ unique permutations into sets with equal numbers of zero responses in the vaccine group. One can think of this partition as being derived from the hypergeometric distribution where we are sampling zeros in the vaccine group. The partition can be written as

$$\left(\begin{array}{c} N \\ n_V \end{array} \right) = \sum_{h=\max(0, n_V - M)}^{\min(n_V, K)} \left(\begin{array}{c} K \\ h \end{array} \right) \left(\begin{array}{c} M \\ n_V - h \end{array} \right) \tag{8}$$

On the right-hand-side of equation 8 the first term in the sum represents the number of ways to select the indices of the zero responses, while the second term represents the number

of ways to select the nonzero responses. Let Q_h be the proportion of the permutation test statistics less than or equal to the observed test statistic among permutations with h zeros in the vaccine group. Specifically,

$$Q_h = \frac{\sum_{j \in \Omega_h} I\left[T_j \leq T_0\right]}{\left(\begin{array}{c} M \\ n_V - h \end{array} \right)} \tag{9}$$

where Ω_h is the set of unique permutations of $\mathbf{Z}_M$ that induce h zeros in the vaccine group. In other words, Ω_h does not include two different permutations of $\mathbf{Z}$ if they only differ within the first $K = N - M$ elements, since those elements are all equal to zero.

The standard calculation groups the $N!$ permutations into $\left(\begin{array}{c} N \\ n_V \end{array} \right)$ sets of unique permutations of $\mathbf{Z}$, and each set has the same number of members. In the case of equation 8, each group with h zeros in the vaccine group does not have the same number of members. The one-sided p-value is a weighted average of the Q_h values:

$$\begin{aligned} \text{p-value} &= \sum_{h=\max(0,n_V-M)}^{\min(n_V,K)} Pr[\text{ a permutation has } h \text{ zeros in the vaccine group}] Q_h \\ &= \sum_{h=\max(0,n_V-M)}^{\min(n_V,K)} \left\{ \frac{\left(\begin{array}{c} K \\ h \end{array} \right) \left(\begin{array}{c} M \\ n_V - h \end{array} \right)}{\left(\begin{array}{c} N \\ n_V \end{array} \right)} \right\} Q_h \\ &= \sum_{h=\max(0,n_V-M)}^{\min(n_V,K)} f(h; K, M, n_V) Q_h \end{aligned} \tag{10}$$

where $f(h; K, M, n_V)$ is the implicitly defined probability mass function of the hypergeometric distribution. Thus to efficiently calculate an exact p-value, we evaluate (10).

Approximate p-value

If m_C and m_V are sufficiently large, then (10) will need to be approximated. Our approach to this approximation is to approximate each Q_h and then calculate the proper weighted combination of $\hat{Q}_h$s.

The key to the approximation is to represent Q_h in a form so that we can use the permutational central limit theorem (PCLT), which we give informally (see Sen [1985] for formal statement).

PCLT: *Consider a permutation where the test statistic is of the form* $\mathcal{T}(\mathbf{S},\mathbf{R}) = \sum S_i R_i$*, and where both of the* $N \times 1$ *vectors of constants,* $\mathbf{S}$ *and* $\mathbf{R}$*, meet some regularity conditions as* N *gets large. Under the assumption that each permutation of* $\mathbf{R}$ *is equally likely,*

$$(N-1)^{-1/2}\left(\frac{\mathcal{T}(\mathbf{S},\mathbf{R}) - N\bar{S}\bar{R}}{\hat{\sigma}_S\hat{\sigma}_R}\right) \dot{\sim} N(0,1) \tag{11}$$

where $\bar{R}, \bar{S}, \hat{\sigma}_R$ *and* $\hat{\sigma}_S$ *are sample means and standard deviations and* $\dot{\sim}$ *denotes approximately distributed for large* N*.*

Let a superscript asterisk denote the sample sizes in the chopped data set (e.g., K^* is the total number of zeroes in the chopped data). When there are h zeroes in the vaccine group in a permutation, this induces h^* zeros in the vaccine group of the chopped data set, where

$$h^* = \begin{cases} h - \lfloor \frac{n_V(K-h)}{n_C} \rfloor & \text{if } \frac{M-n_V+h}{n_C} \geq \frac{n_V-h}{n_V} \\ 0 & \text{if } \frac{M-n_V+h}{n_C} < \frac{n_V-h}{n_V} \end{cases}$$

For both the usual permutation and the rank-based chop lump test, we can represent T_{CL} as a standardized difference in means on the chopped data set, where the responses are scores S_i. For the usual permutation test $S_i = W_i$. There is a slight complication with the rank-based chop-lump test; the rankings are calculated after the chop, so that the scores will change for different permutations. To minimize this problem, we use shifted ranks, i.e., the usual ranks among all $\mathbf{W}$ minus K^*. The shifted ranks give equivalent test results, but the shifted ranks of the non-zero values of $\mathbf{W}$ remain fixed for all values of h. Specifically, for $W_i > 0$ we let $S_i = R_i$ the rank among the $\mathbf{X}$, and when $W_i = 0$ we define the associated S_i values as $S_0 = S_0^{(K^*)} = -(K^*-1)/2$. Thus, in the following, the scores S_i could be either the shifted ranks for the rank based test or the W_i (in which case $S_0^{(K^*)} = 0$ for all K^*) for the usual test.

Letting $\sum^* = \sum_{i=N-N^*+1}^{N}$, we rewrite $T_{CL}(\mathbf{S}, \mathbf{Z})$ for the standardized difference in means tests as

$$\begin{aligned} T_{CL}(\mathbf{S}, \mathbf{Z}) &= \frac{(\sum^* S_i Z_i / n_V^* - \sum^* S_i(1-Z_i)/n_C^*)\left(\frac{1}{n_C^*} + \frac{1}{n_V^*}\right)^{-1/2}}{\hat{\sigma}_{S_{N^*}}} \\ &= \frac{(N^*-1)^{-1/2}\left(\sum^* S_i Z_i - n_V^* \bar{S}_{N^*}\right)}{\hat{\sigma}_{S_{N^*}}\hat{\sigma}_{Z_{N^*}}} \\ &= f_{h^*} + g_{h^*}\mathcal{T}(\mathbf{S}_M, \mathbf{Z}_M) \end{aligned}$$

where f_{h^*} and g_{h^*} are constant through all permutations within Ω_h and are given by

$$f_{h^*} = \frac{h^* S_0^{(K^*)} - n_V^* \bar{S}_{N^*}}{\sqrt{N-1}\hat{\sigma}_{S_{N^*}}\hat{\sigma}_{Z_{N^*}}}$$

and

$$g_{h^*} = \frac{1}{\sqrt{N^*-1}\hat{\sigma}_{S_{N^*}}\hat{\sigma}_{Z_{N^*}}}.$$

Thus, within permutations in Ω_h

$$\begin{aligned} & T_{CL}(\mathbf{S}, \mathbf{Z}) \leq t \\ \Leftrightarrow\ & \mathcal{T}(\mathbf{S}_M, \mathbf{Z}_M) \leq \frac{t - f_{h^*}}{g_{h^*}} \\ \Leftrightarrow\ & \frac{\mathcal{T}(\mathbf{S}_M, \mathbf{Z}_M) - M\bar{S}_M\bar{Z}_M}{(M-1)^{1/2}\hat{\sigma}_{S_M}\hat{\sigma}_{Z_M}} \leq \frac{\frac{t-f_{h^*}}{g_{h^*}} - M\bar{S}_M\bar{Z}_M}{(M-1)^{1/2}\hat{\sigma}_{S_M}\hat{\sigma}_{Z_M}} \end{aligned}$$

Substituting $\bar{Z}_M = \frac{n_V^* - h^*}{M}$ and $\hat{\sigma}_{Z_M} = \sqrt{\frac{(n_V^*-h^*)(M-n_V^*+h^*)}{M(M-1)}}$ and using the PCLT, we approximate Q_h with

$$\hat{Q}_h = \Phi\left(\frac{\frac{T_0 - f_{h^*}}{g_{h^*}} - (n_V^* - h^*)\bar{S}_M}{\hat{\sigma}_{S_M}\left(\frac{(n_V^*-h^*)(M-n_V^*+h^*)}{M}\right)^{1/2}}\right),$$

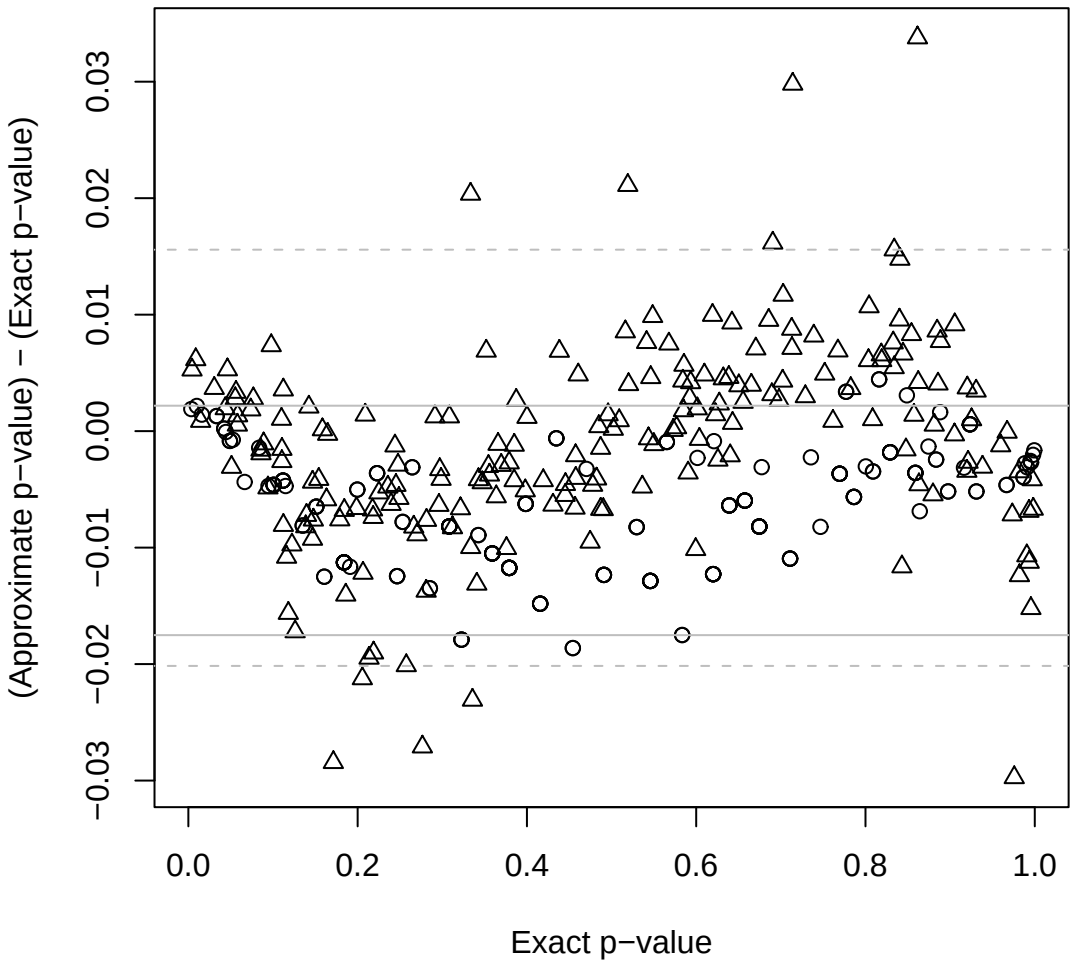
where $\Phi()$ is the standard normal cumulative distribution. We then approximate the p-value by

$$\hat{p} = \sum_{h=\max(0,n_V-M)}^{\min(n_V,K)} f(h; K, M, n_V)\hat{Q}_h. \tag{12}$$

To see how well the approximation performs, we simulate 200 data sets with $N = 100$ and $M = 10$. For this small sample size we can calculate the exact p-values. We randomly

assign $N/2$ of the Z_i values to 1 and the others are 0. We take pseudo-random numbers for the X_i, where $X_i = |X_i^\dagger|$ and $X_i^\dagger \sim N(0,1)$. We plot the bias (approximate p-value minus exact p-value) by the exact p-values in Figure 4, together with the 95% interquantile ranges of the bias. We see that even when M is as small as 10, the approximation does fairly well, with the 95% interquantile range of the bias for the rank tests equal to (-0.0175,0.0022), and the similar statistic for the difference in means tests equal to (-0.0202,0.0156).

Figure 4: Comparison of Asymptotic Approximation and Exact P-values. Data sets have $n_C = n_V = 50$ with 10 total infections. Open circles are Rank Tests and open triangles are Difference in Means Tests



Appendix B: Asymptotics of Rank Tests

In this section we derive asymptotic power expressions for the chop-lump and usual Wilcoxon tests. We also show the asymptotic equivalence of the chop-lump t test and the usual t-test. All results are under local alternatives.

Ordinary Wilcoxon Rank-Sum Test

Assume there are n observations in the vaccine and control arms, respectively. **Take the negatives of viral loads, so that the W_i that are 0 are the maximum values.** The data in the two arms consists of 0s and data:

$$\begin{array}{cccccccc} 0 & 0 & \ldots & 0 & X_{C1} & \ldots & X_{Cm_C} & \\ 0 & 0 & 0 & \ldots & 0 & X_{V1} & \ldots & X_{Vm_V} \end{array}$$

When we combine the control and vaccine observations, the zeroes (remember that the zeroes are the largest ranks) will be ranks $m_C + m_V + 1$ through $2n$. It is intuitively clear that the average rank for the zeroes will be the average of the two numbers $m_C + m_V + 1$ and $2n$. This intuition is made rigorous below.

$$\begin{aligned} \text{average rank for zeroes} &= \frac{(m_C + m_V + 1) + (m_C + m_V + 2) + \ldots + 2n}{2n - m_C - m_V} \\ &= \frac{\sum_{i=1}^{2n} i - \sum_{i=1}^{m_C+m_V} i}{2n - m_C - m_V} \\ &= \frac{(2n)(2n+1)/2 - (m_C + m_V)(m_C + m_V + 1)/2}{2n - m_C - m_V} \\ &= (1/2)\left\{\frac{(2n)^2 + 2n - (m_C + m_V)^2 - (m_C + m_V)}{2n - m_C - m_V}\right\} \\ &= (1/2)\left\{\frac{(2n)^2 - (m_C + m_V)^2}{2n - m_C - m_V} + 1\right\} \\ &= (1/2)\left\{\frac{(2n - m_C - m_V)(2n + m_C + m_V)}{2n - m_C - m_V} + 1\right\} \\ &= (1/2)\{2n + m_C + m_V + 1\}. \end{aligned} \tag{13}$$

Therefore, the sum of the vaccine ranks will be

$$S = \left(\frac{2n + m_C + m_V + 1}{2}\right)(n - m_V) + \sum_{i=1}^{m_V} R_{Vi},$$

where R_{Vi} are the ranks of the m_V nonzero vaccine observations among the m_C+m_V nonzero observations combined across arms (recall that the nonzero observations are negative, so the ranks of the nonzero observations will be $1, 2 \ldots, m_V + m_C$). The standardized Wilcoxon test statistic accounting for ties is

$$Z_{\text{Wil}} = \frac{S - n(2n+1)/2}{\sqrt{\frac{n^2(2n+1)}{12} - \frac{n^2(K^3-K)}{12(2n)(2n-1)}}} = \frac{S - n(2n+1)/2}{\sqrt{\frac{n^2(2n+1)}{12} - \frac{n(K^3-K)}{24(2n-1)}}},$$

where K is the total number of zeroes. We compute the conditional probability of rejection given K. To do this, we first condition additionally on m_V. Conditioning on K and m_V is equivalent to conditioning on m_V and m_C. Substituting our expression for the average rank of the zeroes, we calculate

$$\begin{aligned}
& \Pr(Z_{\text{Wil}} > z_{\alpha/2} \,|\, m_V, m_C) \\
= \;& \Pr\left(\frac{\left(\frac{2n+m_C+m_V+1}{2}\right)(n-m_V) + \sum_{i=1}^{m_V} R_{Vi} - n(2n+1)/2}{\sqrt{\frac{n^2(2n+1)}{12} - \frac{n(K^3-K)}{24(2n-1)}}} > z_{\alpha/2} \,\middle|\, m_V, m_C \right) \\
= \;& \Pr\left(\sum_{i=1}^{m_V} R_{Vi} > z_{\alpha/2}\sqrt{\frac{n^2(2n+1)}{12} - \frac{n(K^3-K)}{24(2n-1)}} + n(2n+1)/2 \right. \\
& \left. - \left(\frac{2n+m_C+m_V+1}{2}\right)(n-m_V) \,\middle|\, m_V, m_C \right) \\
= \;& \Pr\left(\frac{\sum_{i=1}^{m_V} R_{Vi} - m_V(m_C+m_V+1)/2}{\sqrt{\frac{m_V m_C(m_C+m_V+1)}{12}}} > A_n \,|\, m_V, m_C \right), \text{where} \qquad (14)
\end{aligned}$$

$$\begin{aligned}
A_n &= \frac{z_{\alpha/2}\sqrt{\frac{n^2(2n+1)}{12} - \frac{n(K^3-K)}{24(2n-1)}} + \frac{n(2n+1)}{2} - \left(\frac{2n+m_C+m_V+1}{2}\right)(n-m_V) - \frac{m_V(m_C+m_V+1)}{2}}{\sqrt{\frac{m_V m_C(m_C+m_V+1)}{12}}} \\
&= \frac{z_{\alpha/2}\sqrt{\frac{n^2(2n+1)}{12} - \frac{n(K^3-K)}{24(2n-1)}} + n(m_V - m_C)/2}{\sqrt{\frac{m_V m_C(m_C+m_V+1)}{12}}} \qquad (15)
\end{aligned}$$

We can rewrite A_n as

$$\frac{z_{\alpha/2}\sqrt{\frac{n^2(2n+1)}{12} - \frac{n(K^3-K)}{24(2n-1)}}}{\sqrt{\frac{m_V m_C(m_C+m_V+1)}{12}}} - \frac{n z \text{prop}\{2np(1-p)\}^{1/2}/2}{\sqrt{\frac{m_V m_C(m_C+m_V+1)}{12}}},$$

where $z_{\text{prop}} = (m_C - m_V)/\{2np(1-p)\}^{1/2}$ is the Z-score for the test of the proportions infected. Note that large values of z_{prop} indicate that the vaccine works.

Using the fact that the denominator of A_n is asymptotic to $(n^3p^3/6)^{1/2}$, we can show that A_n is asymptotic to

$$z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} - \frac{\sqrt{3(1-p)}z_{\text{prop}}}{p}. \tag{16}$$

Conditioned on m_C and m_V, z_{prop} is a fixed constant. Therefore, conditioned on m_C and m_V, it is as if we are applying a Wilcoxon test to the viral loads in the m_V vaccine infections and m_C control infections, but instead of using critical value $z_{\alpha/2}$, we are using critical value (16).

If we were using a t-test with the same critical value, power would be approximately

$$\begin{aligned} & \Phi\left\{\frac{\Delta}{\sqrt{\sigma^2\left(1/m_V + 1/m_C\right)}} - \left(z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} - \frac{\sqrt{3(1-p)}z_{\text{prop}}}{p}\right)\right\} \\ \approx \quad & \Phi\left\{\frac{\Delta}{\sqrt{2\sigma^2/(np)}} - \left(z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} - \frac{\sqrt{3(1-p)}z_{\text{prop}}}{p}\right)\right\}, \end{aligned} \tag{17}$$

where $\Delta = E[X_V] - E[X_C]$ is positive because we took the negatives of viral loads.

Because we are using the Wilcoxon test instead of the t-test, power is approximately

$$\Phi\left\{\frac{\sqrt{\theta}\Delta}{\sqrt{2\sigma^2/(np)}} - \left(z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} - \frac{\sqrt{3(1-p)}z_{\text{prop}}}{p}\right)\right\}, \tag{18}$$

where θ is the asymptotic relative efficiency of the Wilcoxon test compared to the t-test. For normally distributed outcomes, $\theta \approx .955$.

But remember that (18) is the power conditioned on m_C and m_V. We see that it is a function of m_C and m_V only through z_{prop}. Therefore, (18) is also the power conditioned on z_{prop}. To get the power conditioned on only $m_C + m_V$, we must average over the conditional distribution of z_{prop} given $m_V + m_C$. But z_{prop} is asymptotically independent of $m_V + m_C$, and under a local alternative, Z_{prop} is $\text{N}(d, 1)$. Therefore, power given $m_V + m_C$ is

$$\int_{-\infty}^{\infty} \Phi\left\{ \frac{\sqrt{\theta}\Delta}{\sqrt{2\sigma^2/(np)}} - \left(z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} - \frac{\sqrt{3(1-p)}z}{p} \right) \right\} f_{d,1}(z)dz.$$

$$= \Pr\left\{ Z_2 \leq \frac{\sqrt{\theta}\Delta}{\sqrt{2\sigma^2/(np)}} - \left(z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} - \frac{\sqrt{3(1-p)}Z_1}{p} \right) \right\}, \tag{19}$$

where $Z_1 \sim \mathrm{N}(d, 1)$ and $Z_2 \sim \mathrm{N}(0, 1)$ are independent. We can rewrite (19) as

$$\Pr\left(Z_2 - \frac{\sqrt{3(1-p)}Z_1}{p} \leq \frac{\sqrt{\theta}\Delta}{\sqrt{2\sigma^2/(np)}} - z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} \right).$$

Using the fact that $Z_2 - \{3(1-p)\}^{1/2}Z_1/p$ is normal with mean $-\{3(1-p)\}^{1/2}d/p$ and variance $1 + 3(1-p)/p^2$, we can rewrite power as

$$\Phi\left\{ \frac{\frac{\sqrt{\theta}\Delta}{\sqrt{2\sigma^2/(np)}} - z_{\alpha/2}\sqrt{\frac{1}{p^3} - \frac{(1-p)^3}{p^3}} + \{3(1-p)\}^{1/2}d/p}{\sqrt{1 + 3(1-p)/p^2}} \right\}$$

$$= \Phi\left\{ \frac{\sqrt{\theta}\, p\, \mathrm{E}(Z_{\mathrm{inf}}) + \sqrt{3(1-p)}\, \mathrm{E}(Z_{\mathrm{prop}})}{\sqrt{(1-p)^2 + (1-p) + 1}} - z_{\alpha/2} \right\}. \tag{20}$$

Chop-Lump Wilcoxon

For the chop-lump Wilcoxon test, there are two cases: 1) the leftover zeroes are all in the vaccine arm and 2) the leftover zeroes are all in the control arm.

Case 1: Leftover Zeroes are in the Vaccine Arm ($m_V < m_C$)

Again take the negatives of viral loads, so that the W_i that are 0 are the maximum values.

First condition on m_C and m_V. In Case 1, $m = m_C$, and the number of leftover zeroes after chopping is $m_C - m_V$. Among the $2m$ observations left after chopping, the $m_C - m_V$ zeroes will have ranks $(m_C + m_V + 1), (m_C + m_V + 2), \ldots, 2m$. As in the derivation of (13) the average of these ranks is the average of the smallest and largest rank, namely

$\{(m_C+m_V+1)+(2m)\}/2$. Because $m = m_C$ in Case 1, the average rank is $(3m_C+m_V+1)/2$, and the sum of the vaccine ranks is

$$S = \frac{(3m_C + m_V + 1)(m_C - m_V)}{2} + \sum_{i=1}^{m_V} R_{Vi}, \tag{21}$$

where the R_{Vi} are ranks of the m_V nonzero observations in the vaccine arm among the $m_C + m_V$ nonzero observations in both arms combined.

The Wilcoxon statistic adjusted for ties rejects if

$$\frac{S - m(2m+1)/2}{\sqrt{\frac{m^2(2m+1)}{12} - \frac{m(L^3-L)}{24(2m-1)}}} > c,$$

where L is the number of leftover zeroes, $m_C - m_V$, and c is the critical value, which will be determined later. Substituting (21) for S and standardizing yields

$$\frac{\sum_{i=1}^{m_V} R_{Vi} - m_V(m_C + m_V + 1)/2}{\sqrt{m_C m_V (m_C + m_V + 1)}} > B_n, \text{ where} \tag{22}$$

$B_n =$

$$\frac{c\sqrt{\frac{m}{12}\left\{m(2m+1) - \frac{(L^3-L)}{2(2m-1)}\right\}} + \frac{m(2m+1)}{2} - \frac{(3m_C+m_V+1)(m_C-m_V)}{2} - \frac{m_V(m_C+m_V+1)}{2}}{\sqrt{\frac{m_V m_C (m_C+m_V+1)}{12}}}.$$

It can be shown that $B_n \to c - \sqrt{3(1-p)}z_{\text{prop}}$ as $n \to \infty$. Therefore, asymptotically, (22) is equivalent to using a Wilcoxon test on the m_C and m_V nonzero observations in the control and vaccine arms, but using critical value $c - \sqrt{3(1-p)}z_{\text{prop}}$. Power conditioned on m_C and m_V is therefore approximately

$$(\text{Power} \,|\, m_V, m_C) = \Phi\left\{\frac{\sqrt{\theta}\Delta}{\sqrt{2\sigma^2/(np)}} - c + \sqrt{3(1-p)}\, z_{\text{prop}}\right\}. \tag{23}$$

We now average over the distribution of z_{prop} given $m_V + m_C$, as in the calculation of power for the ordinary Wilcoxon test. This leads to:

$$(\text{Power} \,|\, m_C + m_V) = \Phi\left\{\frac{\sqrt{\theta}\,\text{E}(Z_{\text{inf}}) - c + \sqrt{3(1-p)}\,\text{E}(Z_{\text{prop}})}{\sqrt{1 + 3(1-p)}}\right\}. \tag{24}$$

Now we must determine c such that the test has level α. When $\mathrm{E}(Z_{\mathrm{inf}}) = \mathrm{E}(Z_{\mathrm{prop}}) = 0$, the type 1 error rate is $\Phi\{-c/(1+3(1-p))^{1/2}\}$. To make it a level-$\alpha$ test, we must take $c = (1+3(1-p))^{1/2} z_{\alpha/2}$. With this value of c, power becomes

$$(\mathrm{Power} \,|\, m_C + m_V) = \Phi\left\{ \frac{\sqrt{\theta}\,\mathrm{E}(Z_{\mathrm{inf}}) + \sqrt{3(1-p)}\,\mathrm{E}(Z_{\mathrm{prop}})}{\sqrt{1+3(1-p)}} - z_{\alpha/2} \right\}. \tag{25}$$

Case 2: Leftover Zeroes Are in Control Arm ($m_V > m_C$)

Exactly the same expression for power holds in Case 2 as in Case 1. One way to see this is to write the unstandardized Wilcoxon statistic as the sum of the control ranks and then use the result of the preceding subsection.

Asymptotic Equivalence between chop-lump t-test and ordinary t-test

The chop-lump t-statistic may be written as

$$\frac{\sum_{i=1}^n W_{Ci} - \sum_{i=1}^n W_{Vi}}{\sqrt{2mS_m^2}},$$

where S_m^2 is the pooled sample variance of the $n-m$ largest values in each arm (the values to the right of the chopping point). Assume the null hypothesis is true. We can write the chop-lump t-statistic as

$$\begin{aligned} T &= \left(\frac{\sum_{i=1}^n W_{Ci} - \sum_{i=1}^n W_{Vi}}{\sqrt{2nS_W^2}} \right) \sqrt{\frac{2nS_W^2}{2mS_m^2}} \\ &= \left(\frac{\sum_{i=1}^n W_{Ci} - \sum_{i=1}^n W_{Vi}}{\sqrt{2nS_W^2}} \right) \sqrt{\frac{S_W^2}{\{m/n\}S_m^2}}, \end{aligned} \tag{26}$$

where S_W^2 is the pooled variance of the Ws (zeroes and nonzeroes). Note that m/n converges in probability to the probability of being infected, while S_W^2 and S_m^2 converge to the population variances of the Ws and Xs, respectively. It follows that the square root term on the right converges in probability to a constant, call it ρ. Also, the term on the left in

large parentheses is the ordinary t-statistic applied to the Ws. By Slutsky's theorem, T is asymptotically the same as ρ times the ordinary t-statistic applied to the Ws. Therefore, its asymptotic critical value is 1.96ρ. But rejecting when $T > 1.96\rho$ is the same as rejecting when the ordinary t-statistic on the Ws exceeds 1.96. Therefore, the chop-lump t and ordinary t are asymptotically equivalent; they both reject iff the ordinary t-statistic exceeds 1.96. Note that asymptotically, the proportion of zeroes in the chopped data to the right of the chopping point goes to zero, and thus S_m^2 converges to var$[X]$. If the proportion of zeroes is not negligible, S_m^2 can be substantially larger than var$[X]$. This is especially true if $E[X_C]$ is large, as the zeros are substantial "outliers" relative to the distribution of X. Convergence is slower for larger $E[X_C]$ and slower for small p_V, p_C.