Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Analysis of Methods Used in Trials Containing Many Zeros

Promotor :
Prof. dr. Dan LIN

Promotor :
Mr. TOUFIK ZAHAF

Mari Kassapian

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Analysis of Methods Used in Trials Containing Many Zeros

Promotor :
Prof. dr. Dan LIN

Promotor :
Mr. TOUFIK ZAHAF

Mari Kassapian

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Analysis of Methods Used in Trials Containing Many Zeros

Marie Kassapian

Department of Statistics
University of Hasselt

September 2011

Abstract

In the recent years, many treatments have been developed for the cure of a variety of diseases. Many of these treatments, like the one used for the herpes zoster, are supposed to act in a multilevel way, when administered to patients. They either prevent the virus completely from being expressed or, when this is not possible, since each patient's body may react differently to the same treatment, they ease the pain caused by the disease. Such cases, most of the times, result in data that are consisted of many zeros belonging to those who eventually did not experience the disease. Thus, ways of treating properly this majority of zero scores are necessary, since if these scores are included in the analysis without some attention, they might dilute the real effect of the treatment and the results given might be inaccurate to trust. In this paper, some of the most known approaches, suggested for this issue, are analyzed and compared in terms of their power. Advantages and disadvantages of them are presented as well. Illustrations of some of these methods, with simulated data, are also available as a way of better understanding the concept under which they are working.

Acknowledgements

First of all, I would like to thank my supervisors, Dan Lin and Toufik Zahaf, for their constant help and guidance. Their advices were proved to be priceless for the completion of this project.

I will always be grateful to George for his support throughout the whole master program and for always encouraging me whenever things got tough.

I owe a great thanks to my parents and my sister.

Thanks to all of my friends, especially Georgia, who always believed in me.

Last but not least, i want to thank Marijke Pauwels and George Skiadopoulos for providing technical information.

Contents

1. Introduction

1.1 The Disease

Herpes zoster (HZ), commonly known as *shingles* and also known as *zona*, is a viral disease characterized by a painful skin rash with blisters in a limited area on one side of the body, often in a stripe. The initial infection with Varicella zoster virus (VZV) causes the acute (short-lived) illness chickenpox, which generally occurs in children and young people. Once an episode of chickenpox has resolved, the virus may not be eliminated from the body, but can go on to cause shingles - an illness with much different symptoms - often many years after the initial infection.

Varicella zoster virus can become latent in the nerve cell bodies and less frequently in non-neuronal satellite cells of dorsal root, cranial nerve or autonomic ganglion, without causing any symptoms. Years or decades after a chickenpox infection, the virus may break out of nerve cell bodies and travel down to nerve axons to cause viral infection of the skin in the region of the nerve. The virus may spread from one or more ganglia along nerves of an affected segment and infect the corresponding dermatome (an area of skin supplied by one spinal nerve) causing a painful rash. Although the rash usually heals within two to four weeks, some sufferers experience residual nerve pain for months or years, a condition called postherpetic neuralgia.

The earliest symptoms of herpes zoster, which include headache, fever, and malaise, are nonspecific, and may result in an incorrect diagnosis. These symptoms are commonly followed by sensations of burning pain, itching, hyperesthesia (oversensitivity), or paresthesia (tingling, pricking and/or numbness). The pain may be mild to extreme in the affected dermatome, with sensations that are often described as stinging, tingling, aching, numbing or throbbing, and can be interspersed with quick stabs of agonizing pain. Herpes zoster in children is often painless, but older people are more likely to get zoster as they get older, and then the disease tends to be more severe.

A live vaccine for VZV exists, marketed as Zostavax. In a study of 38,000 older adults (2005), it prevented half of the cases of herpes zoster and reduced the number of cases of postherpetic neuralgia by two-thirds (Oxman, 2005). In October 2007 the vaccine was officially recommended in the U.S. for healthy adults aged 60 and over. Adults also receive an immune boost from contact with children infected with Varicella (chicken pox), a boosting method that prevents about a quarter of herpes zoster cases among unvaccinated adults, but that is becoming less common in the U.S. now that children are routinely vaccinated against Varicella. The shingles vaccination can cut the risk of the severe disease by 55%.

1.2 ZBPI Questionnaire

The intensity, character and duration of postherpetic neuralgia (PHN) vary widely among individuals. Because of the complexity of HZ pain, an accurate and reliable method of pain measurement that captures the magnitude and duration of pain and discomfort is necessary, as the impact of therapeutic or preventive interventions for zoster is needed to be evaluated. In addition to that, it is important to determine the level of pain that interferes with functional status and quality of life to formulate clinically relevant definitions of both acute HZ and PHN (Coplan et al., 2004).

A wide variety of questionnaires have been proposed during the last years, but more or less they all had some limitations, which made it difficult for them to be widely and permanently used. Specifically for herpes zoster, an adaptation of a main pain measure, the Brief Pain Inventory (BPI), has been made in order to create a HZ-specific measure of pain severity that captures discomfort, while also comparing it with other measures of pain as well as functional status and quality of life. Moreover, through this adjusted questionnaire it would be possible for the duration of pain and its severity over time to be measured.

The Zoster Brief Pain Inventory (ZBPI) Questionnaire is based on the BPI, which uses an 11-point Likert scale (0-10) to rate pain in four ways (worst, least, average, now) and pain related interference in seven functional categories (general activity, mood, walking ability, work, relations with others, sleep, enjoyment of life). The reliability of the questions included in this questionnaire, as well as of the questionnaire as a whole, was proved in the end of a study carried out by Coplan et al. (2004).

Examples of other questionnaires that are used for respective reasons are the McGill Pain Questionnaire, the SF-12 and EuroQoL. The McGill Pain Questionnaire is a well established measure for describing the diverse dimensions of pain (McDowell and Newell, 1996). The EuroQoL, or EQ5D, is a validated measure of health-related QoL (Brazier et al., 1993) and consists of five questions and a visual analogue scale.

2. Description of Problems

Nowadays, in many cases it is not possible for a vaccine to prevent completely the case of a disease to occur in a person. For some diseases, attempts are being made in order for vaccines to be discovered, which will thwart their sudden appearance. But when this is not entirely possible, since each individual would react differently to the vaccine, it is of equal importance if an additional privilege of such vaccines were to decrease the patients' pain caused by the disease. One such disease is Herpes Zoster. Relief of acute and chronic HZ pain and discomfort is the main goal of HZ interventions, because pain and other discomfort (e.g. allodynia and intense pruritus) can have a substantial adverse impact on the functional status and quality of life of affected individuals (Coplan et al., 2004).

In randomized, placebo-controlled prevention trials of drugs or vaccines efficacy is typically measured by comparing the rate of occurrence of the disease in the treated group with that in the placebo group. Nevertheless, an intervention may affect both the incidence rate of the disease and its severity. Thus, to evaluate the effects of a preventive intervention in total, it is desirable to take into account both its effect on disease incidence and on the severity (Chang et al., 1994). A combined measure of efficacy, which takes into consideration both of the above mentioned objectives, can be formed by assigning a severity score to each confirmed case of zoster and summing-over all cases to create a burden-of-illness score.

The HZ "Burden-of-Illness (BOI) score" represents the average severity of illness among all patients in the placebo and the vaccine group. It is calculated according to the "modified" scale described by Coplan et al. (2004) as the sum of the HZ "severity-of-illness" scores of all members of the treatment group (vaccine or placebo). If, in addition, it is divided by the total number of subjects in the group it yields the BOI per randomized subject.

Chang et al. (2004) proposed a measure which accounts both for the change of the disease incident rate after vaccination and its severity on those subjects who eventually did experience the disease in a clinical study using the method previously described through the BOI score. This method of the combined comparison of incidence and severity among the treatment groups tends to yield higher power than if these two objectives were tested separately. However, the drawback of such studies is the fact that for a lot of subjects a zero score is expected to be recorded, since the incidence rate of the disease is expected to decrease substantially.

To handle such limitations, Follmann et al. (2009) introduced a new measure of efficacy for the severity and incidence rate, relative in thinking with the one suggested by Chang et al. in the sense that it also makes use of the burden-of-illness score. Its additional privilege is that it tries to handle properly the number of zero scores, since even if there was an important effect of the vaccine, the presence of a massive number of zeros in each group would tend to dilute it.

The testing procedure of Follmann et al. is called “Choplump” test and it has already been implemented in the statistical package R (version 2.10). The main objective of this project is the implementation of this procedure in the statistical software SAS through algorithms and macros. Additional objective is the comparison of the “Choplump” test to the original analysis of the BOI proposed by Chang et al. (1994).

3. Data

Unfortunately, for the analysis real data were not available, since an original clinical trial is still ongoing. Thus, a simulated dataset had to be created and used for the analysis. The simulated dataset was created under specific characteristics and conditions. To begin with, as in most clinical trials, it consists of two treatment groups, placebo and vaccine. The total number of subjects is 16,000 with each group consisting of half the total number, i.e. 8,000 patients. The number of incident cases in each group was calculated under specific assumptions. To be more particular, two main facts were taken into account. First, the fact that the incidence rate is expected to be equal to 0.7% each year throughout the three year study and secondly, the reduction rate of incidents was set to 70%. Hence the cases for each group were computed as follows:

$$m_0 = incidence\ rate * group\ size * years$$

$$m_1 = incidence\ rate * group\ size * risk\ rate * years$$

Under these assumptions, the numbers of incident cases for the placebo and the vaccine group were found to be 168 and 50 respectively. This means that in total there are only 218 values for the BOI score greater than zero. All the rest, non-incident cases, have a zero value for this score. Thus, it is clear that the dataset contains a large number of zeros and the need for ways to correctly handle this issue is obvious.

Regarding the incident cases among the two groups, the simulated scores also have particular characteristics. Since, the new vaccine is expected not only to decrease the number of zoster incidents, but also to lower the Burden-of-Illness score in the incident cases, the scores were generated under a similar concept. More specifically, the BOI (i.e. severity) scores for the vaccine group ranged from 1 to 7, whereas the respective scores for the placebo, expected to be higher due to absence of vaccine delivery, ranged from 4 to 10.

The BOI score for each patient experiencing herpes zoster represented the worst daily score based on the "Zoster Brief Pain Inventory Questionnaire". As explained already, this questionnaire consists of a list of questions related to the pain that a patient feels during the day. This questionnaire has to be filled in every day throughout the whole period during which the patient is followed-up. The worst daily score for each case is recorded, so as to compute the total BOI score in the end of the study by simply adding all the daily scores together. Hence, a patient's final score yields from the summation of 182 daily scores, since this is the period of follow-up that will be considered for the analysis.

4. Description of Methods

4.1 Burden-of-Illness score (Chang et al., 1994)

As it has already been explained in detail, it is important during the evaluation of a vaccine's efficacy that not only the frequency of the incidents, but also the decrease (or not) of the pain due to the disease, to be accounted for. According to Chang's method a severity score is assigned to each individual case experiencing a zoster incident. By adding all the individual scores together, the total score, indicating the magnitude of the burden-of-illness, is computed for each treatment. To test the efficacy of a new intervention, the total BOI scores between the placebo and the vaccine group are compared based on the difference of the BOI scores between the intervention groups. This difference is a measure of the net reduction in morbidity per subject.

Nevertheless, due to the fact that the number of summands that contribute to the calculation of the burden-of-illness score for each group is a random variable, equal to the number of zoster cases in that group, the difference in BOI per treated subject among the treatment groups cannot be analyzed based on ordinary methods.

Assuming a fixed time design, the following parameters are defined. Let n_0 represent the number of subjects randomized to the placebo group, whereas n_1 the number of subjects randomized to the active intervention group, i.e. the vaccine group. For the total number of the study participants, N, it holds that $N = n_0 + n_1$. In a similar way, m_0 stands for the number of zoster cases in the placebo group while m_1 stands for the number of cases in the vaccine group. Again, $M = m_0 + m_V$. The severity scores for the cases, designated W_{01}, $W_{02}, \dots, W_{0n_0}$ for the placebo and W_{11}, $W_{12}, \dots, W_{1n_1}$ for the vaccine group, are assumed to be mutually independent random variables, identically distributed within each group, with respective means μ_0 , μ_1 and variances σ_0^2, σ_1^2. In the end of the study period, the number of cases in both groups can be assumed to follow independent binomial distributions $B(m_0, p_0)$ and $B(m_1, p_1)$ respectively, where p_0 and p_1 are the expected proportions of cases under placebo and active group.

To simplify things, the number of participants is assumed to be equal in each treated group, such as $\mu_0 = \mu_1$. The test statistic T is the difference in burden-of-illness scores per subject:

$$T = \frac{\sum_{i=1}^{n_0} W_{0i}}{m_0} - \frac{\sum_{i=1}^{n_1} W_{1i}}{m_1}$$

It is shown that the test statistic T converges to the standard normal distribution, i.e.

$$\frac{T - E(T)}{\sqrt{Var(T)}} \sim N(0,1)$$

with the two-sided rejection region of the null hypothesis being defined by: $\left|T/\sqrt{\hat{V}(T)}\right| > z_{a/2}$

Furthermore, in the original paper of Chang et al. it can be seen that the power of this test statistic is higher than the power of the statistics used to test either the incidence rate or the severity of the disease separately. To conclude, the sample size needed for such a test is usually smaller than the one that would be appropriate if the trial was designed to detect reduction in incidence, ignoring any differences in severity of pain per case.

4.2 Chop-lump Test (Follmann et al., 2009)

The proposal by Follmann et al. aims at developing a more powerful BOI based test and at obtaining a 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, thus, makes sense to focus only on the side of the distribution's tail where the number reside. Under such a concept, they proposed the "Choplump" test.

The procedure has as follows. A score, say W, equal to 0 is assigned to the non-infected subjects and some disease severity W>0 is assigned to those infected. To test the equality of the lumpy distribution of W between the treated groups, the data are sorted separately in each group and after that the group with the fewer zeros is defined. All the zeros from this group are tossed out and immediately an equal proportion of zeros are thrown out from the other group. In the end, there will be one group with no zeros at all and one group with some zero scores. The scores remained after the chopping are obviously on the right tail of the distribution. A null distribution can be obtained by permutation where the entire dataset is scrambled, a new chopping point is determined and the test statistic is reconstructed.

At this point, it is worth making a small reference to the permutation tests as well as to their usefulness. Permutation tests belong to the variety of resampling methods along with bootstrapping and jackknifing. A permutation test is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. In other words, the method by which treatments are allocated to subjects in an experimental design is mirrored in the analysis of that design. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels. The advantage of permutation tests is that they exist for any test statistic, regardless of whether its distribution is known or not. Thus, one is always free to choose the statistic which best discriminates between hypothesis and alternative and which minimizes losses. On the other hand, an important assumption behind a permutation test, but at the same time its main limitation, is that the observations are exchangeable under the null hypothesis.

Coming back to the "choplump" procedure, similar to the BOI test of Chang et al., a standardized mean difference can be calculated based on the remaining burden-of-illness scores, $W's$, found to the right of the chopping point (for simplicity the notations are the same with those from the Chang et al. test statistic):

$$Chop - lump\ Test\ (CLT) = \frac{\frac{1}{m}(\sum_{i=(n-m)+1}^{n} W_{1(i)} - \sum_{i=(n-m)+1}^{n} W_{0(i)})}{\sqrt{\frac{2S_m^2}{m}}}$$

where n is the number of patients randomized to each group, $m = max\ (m_0, m_1)$ and S_m^2 is the pooled variance based on the m largest $W's$ in each group.

Some details related to the calculation of the p-values for chop-lump tests are worth mentioning, since there are some differences that have to be pointed out, which result from the different datasets that can be used. As mentioned earlier, the choplump test is mainly based on permutation tests. What is, specifically, related to the permutation tests is the calculation of the p-value, which is used for the conclusion regarding the rejection (or not) of the null hypothesis of no difference in the BOI scores among the two experimental groups. In cases, where there is a small number of subjects in the study, the computation of an exact p-value is feasible. This is done by accounting for all the possible permutations that can be performed in the dataset. The problem arises, when the number of the possible permutations is large. It, then, becomes computationally very intensive for an exact p-value to be computed. Thus, it is important to calculate it, based on a sample from all the permutations, which is randomly selected.

Again, let $N = n_0 + n_1$ and $M = m_0 + m_1$ represent the total number of subjects in the study and the total number of zoster cases, respectively. As can be seen, both numbers can be partitioned into the relative numbers for the placebo and the vaccine group. Also, let $K = k_0 + k_1$ stand for the total number of cases where an incident of zoster was not recorded. In other words, K is the total number of zero scores in both groups. Consider that the data can be represented by the two vectors $W = [W_1, W_2, \dots, W_N]$ and $Z = [Z_1, Z_2, \dots, Z_N]$, where Z is the group membership indicator. The appropriate procedure is as follows. The indices $j = 1,2, \dots, N$ are ordered first by W_j and then by Z_j within tied W_j values, so that $Z_1, Z_2, \dots, Z_{k_0}$ are zeros and $Z_{k_0+1}, Z_{k_0+1}, \dots, Z_K$ are ones. Let W_a and Z_a be the last a values of $W$and Z, respectively. Let 0_a and 1_a be vectors of zero or one of length a, where $a = 0$ denotes no vector.

❖ ***Case 1:*** **P-value based on all permutations**

In the usual permutation test, a test statistic T is defined, which is a function of W and Z. Let T_0 be the test statistic evaluated at the original data, and T_i be the test statistic evaluated at the i^{th} permutation of the values of Z. If lower values of the test statistic are more extreme, then:

$$(one - sided)\, p - value \; = \frac{\sum_{i=1}^{N!} I\{T_i \leq T_0\}}{N!} \qquad (1)$$

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

❖ ***Case 2:*** **P-value based on a specific number of permutations**

In cases, where it is impossible to consider the whole number of permutations, the calculation of the p-value can be based on a Permutational Central Limit Theorem (PCLT). 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, i.e.

$$Q_h \; = \frac{\sum_{i \in Q_h} I\{T_i \leq T_0\}}{\binom{M}{n_1 - h}}$$

where Q_h is the set of unique permutations of Z_M that induce h zeros in the vaccine group. This means that Q_h does not include two different permutations of Z if they only differ within the first $K \; = \; N - M$ elements, since those elements are all equal to zero (Follmann et al., 2009). The main idea is to represent Q_h in a way such that a PCLT can be used to approximate the p-value (for an explanation of the PCLT we refer to Follmann et al. (2009)). The standard calculation groups the $N!$ permutations into $\binom{N}{n_1} = \sum_{h=\max(0, n_1 - M)}^{\min(n_1, K)} \binom{K}{h} \binom{M}{n_1 - h}$ sets of unique permutations of Z, and each set has the same number of members. The one-sided p-value is a weighted average of the Q_h values:

$$p - \widehat{value} = \sum_{h=\max(0, n_1 - M)}^{\min(n_1, K)} \Pr\,(a\ permutation\ has\ h\ zeros\ in\ vaccine\ group)\, \hat{Q}_h =$$

$$= \sum_{h=\max\,(0, n_1 - M)}^{\min\,(n_1, K)} \left\{ \frac{\binom{K}{h} \binom{M}{n_1 - h}}{\binom{N}{n_1}} \right\} \hat{Q}_h = \sum_{h=\max\,(0, n_1 - M)}^{\min\,(n_1, K)} f(h; K, M, n_1) \hat{Q}_h \qquad (2)$$

where $f(h; K, M, N_1)$ is the probability mass function of the hypergeometric distribution.

4.3 Chop-lump Test - Algorithm in SAS

The implementation of the "Choplump" test was based on two different approaches. The first one covers the case where the number of permutations, possible in a dataset, is relatively small, whereas the second one is more appropriate, when large datasets are present. The difference lies, basically, on the computation of the p-value that is needed for the conclusion of the hypothesis tested. In the first case, the complete number of permutations is taken into account to calculate the p-value, while in the latter one, only a specific number of permutations is considered.

The main steps of the algorithm - for the first case - are the following:

- ***STEP 1:*** Define the main parameters, i.e. $N, n_0, n_1, M, m_0, m_1, K, k_0, k_1$.
- ***STEP 2:*** Chop the dataset.
- ***STEP 3:*** Apply the Wilcoxon function to the chopped dataset to compute the test statistic T_0 of the original data.
- ***STEP 4:*** Create the matrix whose rows represent all possible permutations of the group membership indicators Z_j.
- ***STEP 5:*** Compute all test statistics T_i. Each test statistic is based on the dataset consisted of the vector W of the scores and the vector Z of the group indicators. For each T_i, vector Z is replaced by the i^{th} row of the matrix from step 4.
- ***STEP 6:*** Calculate the p-value based on equation (1).

The main steps of the algorithm - for the second case - are the following:

- ***STEP 1:*** Create function Qh, which calculates the product of the Pr[a random permutation will have h zeros in the vaccine group] and the proportion Qh of the permutation test statistics less than or equal to the observed test statistic among permutations with h zeros in the vaccine group.
- ***STEP 2:*** Define the main parameters, i.e. $N, n_0, n_1, M, m_0, m_1, K, k_0, k_1$.
- ***STEP 3:*** Chop the dataset.
- ***STEP 4:*** Apply the TDiM function – which is based on PCL theorem - to the chopped dataset to compute the test statistic T_0 of the original data.
- ***STEP 5:*** Create a vector consisting of all possible values for h (=number of zeros in the vaccine group).
- ***STEP 6:*** Apply function Qh on the vector created in step 5.
- ***STEP 7:*** Calculate the p-value based on equation (2).

For an illustration of both algorithms, we refer to section 5.2.

4.4 Comparison of Choplump test and Chang's BOI measure

Comparison of the choplump test with the test statistic proposed by Chang et al. was thought to be of great interest, since both are methods used to handle a particular issue in clinical trials, i.e. the presence of many zeros. On the other hand, each one uses a different approach and this aspect of them might be considered important.

Initially, both test statistics were applied on the dataset used throughout the whole study and the results are presented in terms of the p-values. As will be seen later, both tests reject the null hypothesis of the equality of means with very significant p-values. Nevertheless, this was deemed inadequate. Hence, a more thorough approach was adopted. The probabilities of type I and type II error were estimated for a number of scenarios of the sample size and the expected risk reduction. In particular, five scenarios were assumed for the sample size (N = 1,000 , 3,000 , 5,000 , 10,000 and 20,000) and for each one of them, six different cases that were combinations of the risk reduction (RR = 30% , 50% and 70%) and the severity reduction (SR = yes, no), meaning the reduction in the burden-of-illness scores (table 6). Also, values for the incidence rate based on a $Normal(0.007, \sigma^2)$ distribution - by varying the values for σ^2 and keeping the integers- were considered, but the results did not differ much, thus this approach was not eventually adopted.

The type I error probability (α% level of significance) is an indicator of how many times will the test statistic reject the null hypothesis, when this hypothesis actually holds. On the contrary, type II error probability (β) can show how many times will the test fail to reject the null, when the alternative hypothesis is true. Both need to remain at low levels. The latter probability is usually better expressed in terms of the power, i.e. the quantity $1 - \beta$. Decreasing sample size as well as decreasing risk reduction are supposed to lower the power of the test, while increasing values for both quantities are expected to attribute a higher power to it.

For this attempt to be achieved, datasets were simulated under the null and under the alternative hypothesis. Under the null hypothesis, where no risk reduction is assumed, one thousand datasets were simulated for each case of the sample size. Under the alternative hypothesis the assumptions under investigation concern the reduction of the risk and the presence or absence of the severity reduction. Thus, one thousand datasets were simulated for each combination of the sample size, the risk reduction and the severity reduction. For each sample size there were six different alternative hypotheses compared to the same null. On each set of the thousand datasets the two test statistics were applied. Based on the resulting p-values of the datasets simulated under the null hypothesis, the probability of type I error was estimated for each setting. Similarly, the p-values from the alternative hypotheses' datasets were used to estimate the probability of type II error and the power of each test for the different combinations of the population size and the risk reduction.

4.5 Other Studies

A wide variety of studies from other scientific fields have carried out their analyses based on the same thinking of Follmann et al. What is implied by this is that in cases where there are a lot of zeros that should be treated in a specific way so as to make the analysis easier, a widely acceptable method is to exclude these zeros according, of course, to some principals. Some examples of such studies will be presented in the following sub-sections.

4.5.1 Two - part permutation tests for DNA methylation and microarray data (Neuhäuser et al., 2005)

It is widely known that during the last decades, cancer has evolved to a striking disease attacking more and more people, and in most of the cases it has fatal consequences on the infecteds. For this reason, many studies keep taking place in order to discover ways to face cancer in its different forms. One of the latest discoveries is that DNA methylation analysis promises to become a powerful tool in cancer diagnosis.

When the tested region is not or only partially methylated the result is negative (undetectable methylation) and it is then assigned a zero value. In contrast, samples that show methylation will have a value greater than zero. In microarray data studies a common method for small or negative expression levels is to clip them off so as to be equal to an arbitrarily cut-off value. Due to the truncation, there are two different types of data, truncated values and original observations. But, since the truncated values are not just another point on the continuum of possible values, it would be inappropriate to use a standard statistical method that would treat all values equally (Neuhäuser et al., 2005). For such cases, where such a data structure appears, Lachenbruch introduced the "two-part models" (Lachenbruch, 1976; 2001; 2002). This type of model is represented by a test statistic, which is the sum of two squared statistics. The first one compares the proportion of zeros, i.e. the proportion of truncated values in the data, whereas the second compares the positive values.

Let n_1 and n_2 be the numbers of independent observations regarding one gene, for two groups to be compared. Furthermore, let m_1 and m_2, respectively, represent the observed numbers of truncated values (i.e. null values in the case of methylation data). To compare m_1 and m_2, Lachenbruch (1976) used the following statistic:

$$B^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1-\hat{p})\frac{n_1+n_2}{n_1 n_2}} \text{ , where } \hat{p}_1 = {}^{m_1}\!/_{n_1} \text{ , } \hat{p}_2 = {}^{m_2}\!/_{n_2} \text{ and } \hat{p} = {}^{(m_1 + m_2)}\!/_{(n_1 + n_2)} \, .$$

Under the null hypothesis the proportions of zeros truncated are the same between the two groups and B^2 is asymptotically χ^2-distributed with 1 degree of freedom (d.f.). In the two extreme cases where the two groups contain only zero values or no zero values at all then the statistic B^2 cannot be well defined and is set equal to zero ($B^2 = 0$).

Concerning the second factor, a Wilcoxon rank sum test can be used, since non-parametric tests are more suitable for microarray data, because they are usually non-normally distributed. Thus, the statistic based on the positive expression levels is formulated as:

$$W = \frac{RS - [(n_1 - m_1)(n_1 - m_1 + n_2 - m_2 + 1)/2]}{\sqrt{(n_1 - m_2)(n_2 - m_2)(n_1 - m_1 + n_2 - m_2 + 1)/12}}$$

where RS is the sum of the ranks in group 1. In the extreme case, where there are only zero values in at least one of the two groups, the above statistic is set equal to zero ($W = 0$).

Finally, the test statistic for the "two-part model" is constructed as $X^2 = B^2 + W^2$. The null hypothesis, obviously, implies no difference between the two groups. Because B^2 and W are independent and asymptotically normal, the sum of the squared statistics is asymptotically and χ^2-distributed with 2 d.f. This test, called two-part permutation test here, is a permutation test based on the sum statistic X^2. It is carried out by permuting the group labels for the whole sample. When performing this two-part permutation test the exact permutation distribution of X^2 is determined by generating all possible permutations. When the number of permutations is large the p-value can be approximated by using a random sample of all possible permutations.

4.5.2 Hypothesis Tests for Point-Mass Mixture Data with Application to Omics Data with Many Zero Values (Taylor & Pollard, 2009)

Another method to handle data including many zero values was proposed by Taylor & Pollard (2009). According to this paper, biological studies often generate point-mass mixture data composed of a continuous component plus a point-mass – usually at zero. This point-mass can reflect true zeros, such as an absent compound, or truncated values. Truncation results from either a detection limit of the assay or a lower bound on meaningful signal set by the researcher.

Point-mass mixture data are characterized by the proportion of zeros and the distribution of the continuous component, which is assumed to have its support on the non-negative real numbers. In studies, where interest lies in differences between two (or more) experimental groups, if the data in each experimental group is distributed as a point-mass mixture, a difference in means between the experimental groups can result from a difference in the proportion of zeros, a difference in the mean of the continuous component, or both. Taylor and Pollard (2009) specifically point out that, standard statistical methods can be used to focus on one of these effects at a time. Tests with sufficiently loose assumptions about the data distribution may be appropriate, but still fail to distinguish the contributions of the two mixture components to differences between experimental groups. For these reasons it is of great importance to employ statistical tests that account for the separate contributions of each component of a point-mass mixture.

For the notation, consider a two-sample problem with m samples from experimental group 1 and n samples from experimental group 2. Let $x_1, \dots, x_m$ and $y_1, \dots, y_n$ be independent, identically-distributed bivariate observations of $X = (Z, D)$ from group 1 and $Y = (W, H)$ from group 2, respectively. Denote an observation of X by $x = (z, d)$ and an observation of Y by $y = (w, h)$, where z (respectively w) is a non-negative real number and d (respectively h) is an indicator variable with values d (respectively h) =1 if z (respectively w) > 0 and d (respectively h) =0 if z (respectively w) = 0. Then, the probability distribution of x is $f\ (z, d) = p_1^{1-d}\{(1 - p_1)g(z, \mu_1)\}^d$ where $g(z, \mu_1)$ is a parametric density with mean μ_1. Similarly, $f\ (w, h) = p_2^{1-h}\{(1 - p_2)g(w, \mu_2)\}^h$ where $g(w, \mu_2)$ is a parametric density with mean μ_2.
Based on the usual LRT statistic:

$$\Lambda(x, y) = \frac{\prod_{i=1}^{n_1} \hat{p}^{1-d_i}\{(1-\hat{p})g(z, \hat{\mu})\}^{d_i} \prod_{i=1}^{n_2} \hat{p}^{1-h_j}\{(1-\hat{p})g(w, \hat{\mu})\}^{h_j}}{\prod_{i=1}^{n_1} \hat{p}_1{}^{1-d_i}\{(1-\hat{p}_1)g(z, \hat{\mu}_1)\}^{d_i} \prod_{i=1}^{n_2} \hat{p}_2{}^{1-h_j}\{(1-\hat{p}_2)g(w, \hat{\mu}_2)\}^{h_j}}$$

they propose a modification of it by replacing the parametric distributions $g(., \mu)$ with empirical distributions to avoid distributional assumptions. More particularly, their suggestion is to replace $\prod_i^m \left(\frac{g(z,\hat{\mu})}{g(z,\hat{\mu}_1)}\right)^{d_i} \prod_j^n \left(\frac{g(w,\hat{\mu})}{g(w,\hat{\mu}_2)}\right)^{h_j}$ in the above equation with $R(\mu) = \prod_{i=1}^{m_c} m_c r_i \prod_{j=1}^{n_c} n_c q_j$, where m_c and n_c are the number of x and y observations greater than zero and computation of r_i and q_j is based on Lagrange's multiplier method. Since $-2logR(\mu)$ is asymptotically distributed as χ_1^2, the two-part empirical LRT is asymptotically χ_2^2.

4.6 Software

For the exploratory analysis of the data and the plots, as well as for the programming of the macros, the statistical package SAS (version 9.2) was used, whereas the simulation of the datasets, used in the power analysis, was implemented in R (version 2.13.1).

5. Results

5.1 Exploratory Data Analysis

The frequencies of the incidents among the groups as well as the main descriptive statistics for the variable representing the burden-of-illness score were calculated. In the table that follows *(table 1)* the frequencies are presented by group.

Table 1: Frequencies of incidents

	Frequency	**Cumulative Frequency**	**Percent (%)**
Z=0	168	168	77.06
Z=1	50	218	22.94

Regarding the summary statistics of the BOI score, these are presented in table 2. The table is partitioned into two parts. The first one contains the descriptive statistics for BOI based on the whole population, while the second one is based only on those individuals for whom a zoster incident was officially confirmed.

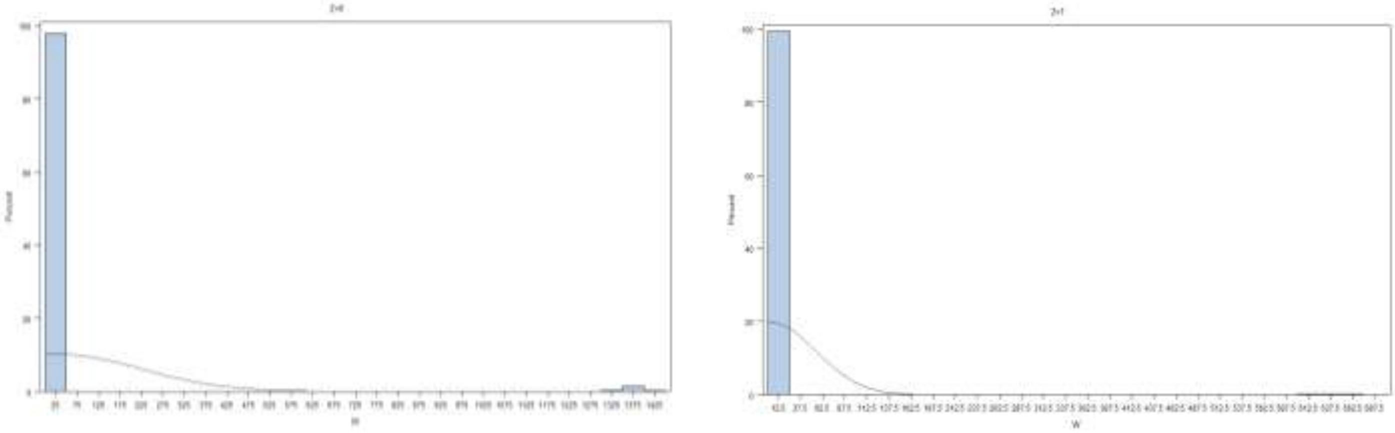
Table 2: Descriptive Statistics for BOI

	N	**Mean**	**Std. Deviation**	**Median**	**Minimum**	**Maximum**
			All cases (W ≥ 0)			
Placebo (Z=0)	8,000	28.69	195.92	0	0	1431
Vaccine (Z=1)	8,000	4.01	50.58	0	0	690
			Zoster cases only (W > 0)			
Placebo (Z=0)	168	1366.20	21.60	1366	1320	1431
Vaccine (Z=1)	50	641.54	21.02	641	597	690

As can be seen in the upper table the mean and the median are far apart, something which does not hold for the lower table. This indicates that for the first case the data will most probably not follow a normal distribution, whereas the opposite will hold for the second case. Normality tests to confirm these assumptions were carried out and will be immediately presented.

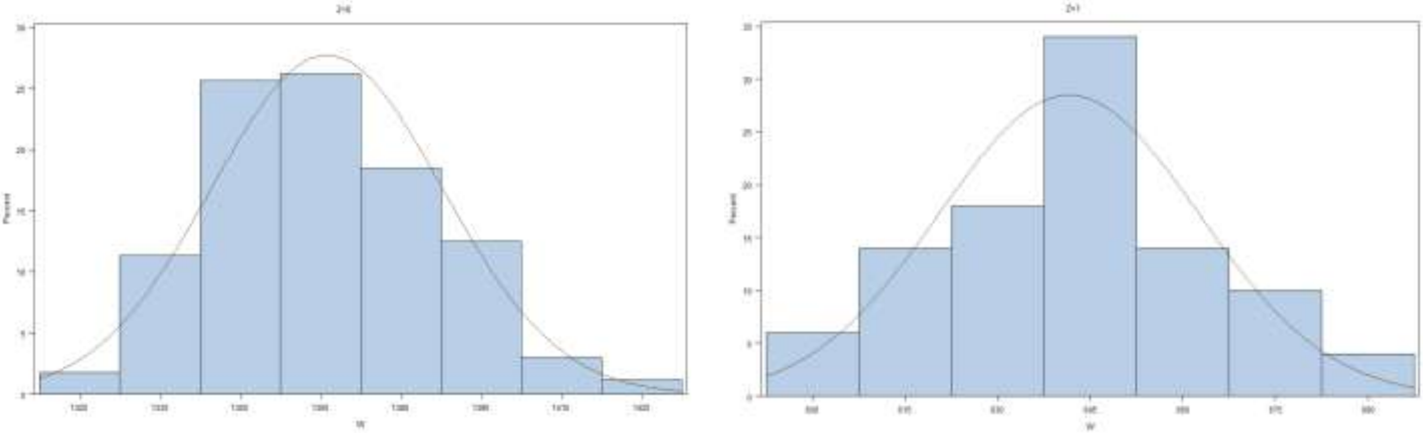
Under the null hypothesis, which assumes that normality of the data holds, again two alternative datasets were considered. The first was the dataset containing the incident as well as the non-incident cases and the second, the one containing only the incident cases. As was expected, since the distribution of the data is zero inflated, the normality test of Kolmogorov-Smirnov yielded highly significant p-values. For both the placebo and the vaccine group, the p-value was less than 0.001, indicating a strong violation of the data normality. On the other hand, concerning the dataset based only on the zoster cases, the respective p-values were equal to 0.128 (placebo) and 0.15 (vaccine), concluding that the normality assumption, indeed, holds.

The following figures show the histograms for each type of dataset for both treatment groups *(figures 1-2)*.



(1a) ***(1b)***

Figure 1: Histograms for placebo and vaccine group including all cases (zoster & non-zoster)



(2a) ***(2b)***

Figure 2: Histograms for placebo and vaccine group including only zoster cases

5.2 Illustration of the Algorithms

The two different types of algorithms, consisted of SAS macros, were applied on two simulated datasets respectively. For the first algorithm, which calculates an exact p-value based on all permutations, a relatively small dataset was used. Specifically, ten patients were considered in total that were equally distributed among the two treated groups. For the placebo group, four patients were considered to experience a zoster incident whereas the respective number of patients in the vaccine group was one. The number of permutations based on ten patients and five incidents is 252. Although, the dataset may seem too small, even just a duplication of its size and its number of cases would yield a number of permutations equal to 184,756 making it computationally prohibitive for an exact p-value to be calculated.

In order for the structure of the dataset to be clearer, table 3 was created.

Table 3: Structure of dataset

Patient ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
W=score	0	0	0	0	0	1326	1369	1387	1374	650
Z=group	0	0	0	0	1	0	0	0	0	1

As can be concluded from the above table the dataset consists of two rows (which are converted into columns when used in the statistical software); the first indicating the sum of the worst daily scores during the 182 days of the follow-up period for each patient and the second indicating the group membership (0 for placebo group and 1 for vaccine group).

Under the null hypothesis of no difference in the BOI scores between the two groups, the SAS macro *%choplumpExact* was applied on this dataset (all macros are available in the Appendix). Considering a level of significance equal to 0.05, the resulting p-value was equal to 0.047, implying that, indeed, the groups differ in two ways. First, they differ in the number of incidents recorded, since the placebo group has recorded more zoster cases. Secondly, they also differ in the severity scores since the placebo group, clearly, has on average higher BOI scores that he vaccine group.

The second dataset, as has already been explained, is based on a population size of 16,000 individuals that are equally divided into two groups. From the patients belonging to the placebo group, 168 of them experienced a zoster incident, while the number of confirmed cases recorded for the active treatment group was equal to 50.

The SAS macro *%choplumpApprox* was applied on this dataset. Based on the steps of the algorithm, not all of the possible permutations were considered. In particular, only 218 permutations out of the total number were taken into account for the computation of the p-value of the hypothesis testing. It is easy to extract that this number is equal to the total number of cases in both groups. One can think of each permutation – out of the chosen ones – as the number of incident cases (or the number of non-incident cases) in the vaccine group. To be more specific, had all cases occurred in the vaccine group, there would be 218 scores greater than zero and no zero score in this group. On the other hand, in the second extreme case where all the cases would have been recorded for the placebo group, there would not be any non-zero score and 218 zero scores in the vaccine group. A better image of the distribution of the scores among the groups can be obtained by the following table *(table 4).*

Table 4: Distribution of (non-)incident cases in vaccine group

	Number of incident cases in vaccine group	**Number of non-incident cases in vaccine group**
All 218 cases in placebo group	0	218
All 218 cases in vaccine group	218	0

The cases described above represent the two extreme cases. All the other cases lie in between these two.

The p-value obtained was equal to $2.72 * 10^{-31}$, resolving to a very small number. As a consequence, it can be, certainly, implied that the null hypothesis is not valid and, thus, has to be rejected at a 5% of significance level. From this point of view, the two groups show a different behavior not only in terms of the incident rate, but also in terms of the severity scoring with the group receiving the new treatment, i.e. the vaccine, operating the better behavior of the two.

5.3 Comparison of Choplump test and Chang's BOI measure – A Simulation Study

At this point, it would be of interest to compare the two basic methods that have been suggested, so far, for handling data with many zeros. These are the method based on the Choplump test of Follmann et al. (2009) and the method proposed by Chang et al. (1994) based on the burden-of-illness scores. (From now on, reference to this method will be acknowledged by "Chang method")

For Chang's method, table 5 was created. The first half of the table shows the basic quantities necessary for the implementation of the test. The second half of it indicates the resulting values of the test statistics for each hypothesis testing and the respective yielded p-values.

Table 5: Tests between the two groups

Treatment Group	**Placebo**	**Vaccine**
Number of individuals	8,000	8,000
Number of cases	168	50
Sum of severity scores	229,522	32,077
Mean score (per case)	1366.20	641.54
Variance of scores	466.73	441.76
Test	**Statistic**	**p-value**
Incidence Rate	63.87	<0.0001
Severity score per case	209.49	<0.0001
Burden-of-illness score	11.22	0.00*

(*): Not exactly zero, but a value very close to zero

The incidence rates of the groups were compared by a chi-square test. The proportion of incidents in the placebo group (77.06%) was compared with the relative proportion in the vaccine group (22.94%) under the null hypothesis of their equality. The hypothesis testing for the equality of the severity scores per case was carried out based on a t-test, which compared the mean severity score in each group. The null hypothesis stood, again, for the equality of the group means per case. Since the variances of the severity scores in the two groups were found to be equal (p-value=0.85), the statistic calculated based on the Pooled method was taken into consideration. Last but not least, the test statistic proposed for this method (Section 4.2) was implemented for the testing of the null hypothesis of equality of the BOI scores among the treated groups. All tests were performed using a 5% level of significance. Based on the p-values, shown in table 5, all null hypotheses are rejected.

For the difference in the groups to become more obvious, the Area Under the Curve was constructed, based on the mean daily scores plotted against time. The mean BOI score for each day was calculated by the sum of all scores reported (in each group) for the "worst pain" question in the ZBPI questionnaire, divided by the total number of patients (in that group). Figure 3 shows a strong difference between the placebo and the vaccine group, with the latter one yielding mean daily scores much lower than those in the former and therefore presenting a smaller area under the curve. Thus, the efficacy of the active treatment is unquestionable.

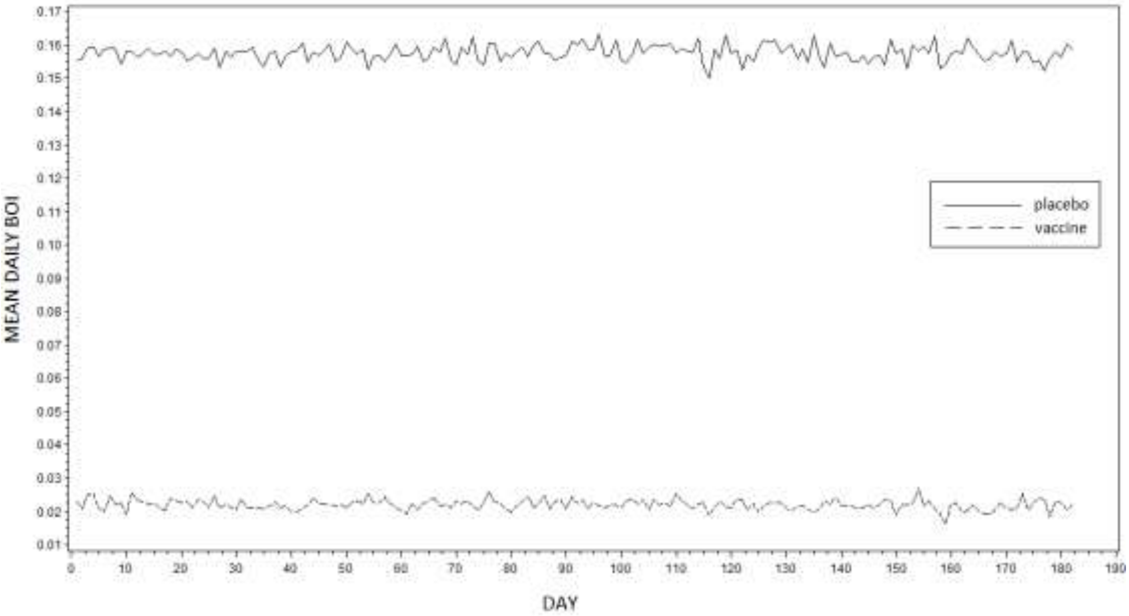


Figure 3: Area Under the Curve for BOI scores of the groups

After the implementation of the test for the equality of the burden-of-illness scores in the two groups based on the two different methods presented in Section 4, the following conclusion can be drawn. Both methods reject the null hypothesis, thus rejecting the equality of the BOI scores in the groups, but with the difference that the Chang method cannot compute values for the p-value that are very small, thus it assigns a value equal to zero.

For this reason, power analysis was carried out in order to investigate whether the Choplump test is more powerful or whether this assumption holds for the alternative method. For this procedure to be carried out, different settings for the sample size, the risk reduction and the severity reduction were adopted and are shown in table 6. The notation of the parameters is the same with the one used for the explanation of the test statistics.

Table 6: Scenarios assumed for the power analysis

Scenario	**N**	**n_0**	**n_1**	**Risk Reduction**	**Severity Reduction**	**M**	**m_0**	**m_1**
1st	1,000	500	500	30%	Yes	19	11	8
2nd					No			
3rd				50%	Yes	17		6
4th					No			
5th				70%	Yes	14		3
6th					No			
7th	3,000	1,500	1,500	30%	Yes	55	32	23
8th					No			
9th				50%	Yes	48		16
10th					No			
11th				70%	Yes	42		10
12th					No			
13th	5,000	2,500	2,500	30%	Yes	89	52	37
14th					No			
15th				50%	Yes	78		26
16th					No			
17th				70%	Yes	68		16
18th					No			
19th	10,000	5,000	5,000	30%	Yes	179	105	74
20th					No			
21st				50%	Yes	158		53
22nd					No			
23rd				70%	Yes	137		32
24th					No			
25th	20,000	10,000	10,000	30%	Yes	357	210	147
26th					No			
27th				50%	Yes	315		105
28th					No			
29th				70%	Yes	273		63
30th					No			

In figure 4, the boxplots of the scores belonging to the null and each alternative hypothesis, for both groups, are presented. As expected, the boxplots under H_0 differ only slightly among the two treated groups, since no risk reduction is supposed to occur and also the range of scores used was the same for all patients, i.e. from 1 to 10. Concerning the boxplots, under the alternatives $H_A^{(2)}$, $H_A^{(4)}$ and $H_A^{(6)}$ no substantial differences can be spotted, since no reduction in the severity is assumed, thus, the score ranges used for both groups were again the same (1-10). In contrast, large differences can be detected for the alternatives considering severity reduction, i.e. $H_A^{(1)}$, $H_A^{(3)}$ and $H_A^{(5)}$. For each of them, scores from 4 to 10 were used for the placebo group, while the ranges 3-9, 2-8 and 1-7, respectively for each alternative hypothesis, included the scores of the vaccine group. The most significant thing that can be extracted from the graph is that the score differences increase considerably with the increase in the risk reduction.

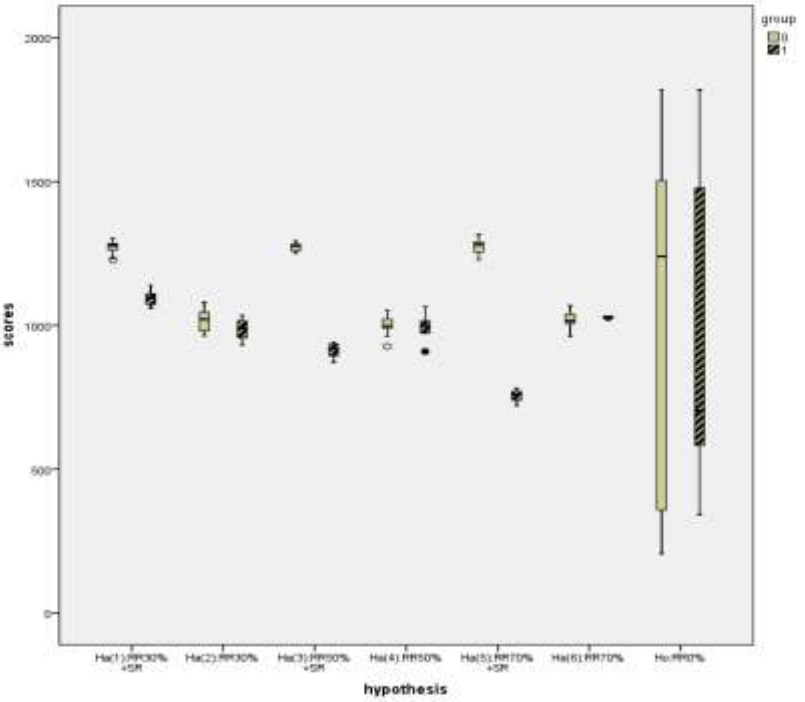


Figure 4: Boxplots of the patients' scores under the different hypotheses (N=10,000)

In order to better understand the trend of the p-values for each case, the main descriptive statistics were calculated under each scenario and are presented in the table that follows (*table 7*). What can be obtained from this table is that, regarding the H_0 and no matter the sample size, the two test statistics do not differ much. Nevertheless, as the sample size increases, the Choplump test gives much smaller p-values. The same behavior holds for this test, when considering the same sample size, but different magnitudes of the reduction in the risk. This implies that, the more the risk declines the more substantial is the decrease of the p-values. For a graphical presentation of the trend of the p-values, the reader can refer to figure 5.

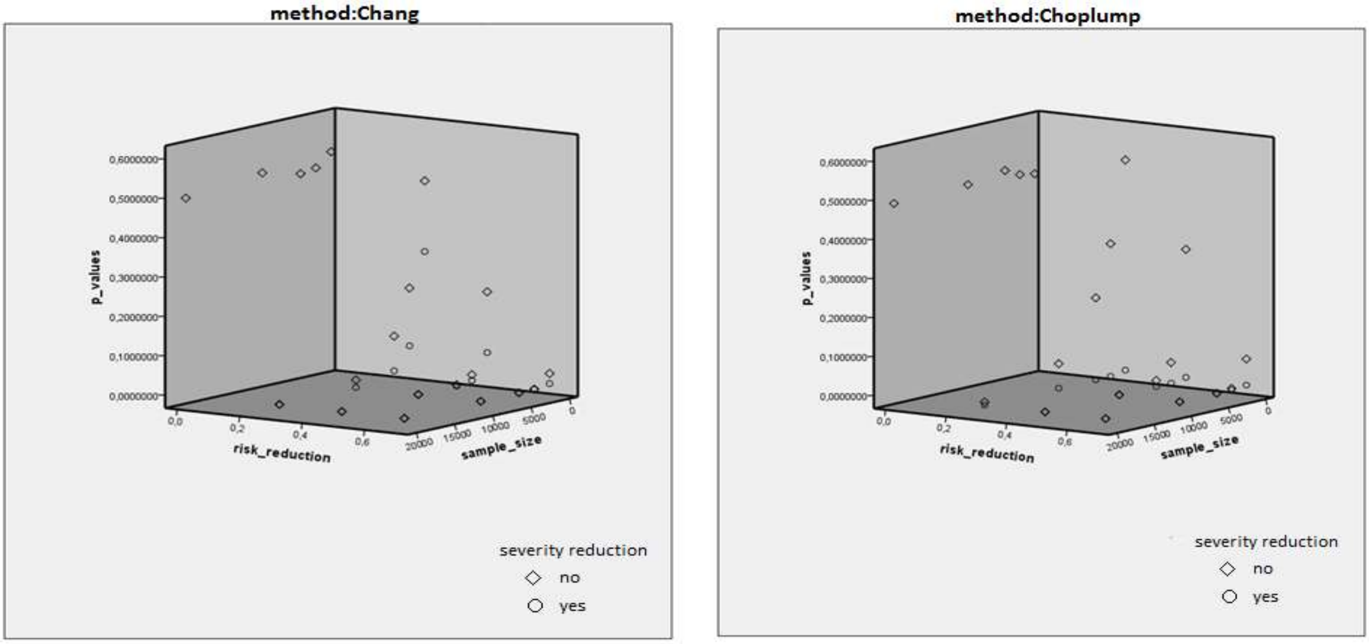
Table 7: Descriptive statistics for the resulted p-values

Case for N	Hypothesis	Method	Mean	Variance	Minimum	Maximum
1st	H_0^*	Chang	0.534	0.033	0.010	0.795
		Choplump	0.484	0.044	0.006	0.800
	$H_A^{(1)*}$	Chang	0.307	0.0001	0.278	0.337
		Choplump	0.0072	$4*10^{-10}$	0.0071	0.0073
	$H_A^{(2)*}$	Chang	0.486	0.0005	0.422	0.568
		Choplump	0.546	0.058	0.0337	0.989
	$H_A^{(3)*}$	Chang	0.068	10^{-5}	0.056	0.079
		Choplump	0.006	$2*10^{-10}$	0.0059	0.0061
	$H_A^{(4)*}$	Chang	0.222	0.0002	0.185	0.268
		Choplump	0.334	0.038	0.007	0.946
	$H_A^{(5)*}$	Chang	0.006	10^{-7}	0.004	0.007
		Choplump	0.0036	$7*10^{-11}$	0.0035	0.0037
	$H_A^{(6)*}$	Chang	0.03	$6*10^{-6}$	0.025	0.040
		Choplump	0.071	0.003	0.0036	0.289
2nd	H_0	Chang	0.502	0.038	0.013	0.768
		Choplump	0.491	0.042	0.050	0.800
	$H_A^{(1)}$	Chang	0.076	10^{-5}	0.064	0.088
		Choplump	$3*10^{-6}$	10^{-18}	$2.8*10^{-6}$	$2.83*10^{-6}$
	$H_A^{(2)}$	Chang	0.222	0.0002	0.179	0.271
		Choplump	0.340	0.044	0.008	0.992
	$H_A^{(3)}$	Chang	0.005	10^{-7}	0.004	0.006
		Choplump	10^{-6}	$3*10^{-17}$	$1.4*10^{-6}$	$1.5*10^{-6}$
	$H_A^{(4)}$	Chang	0.021	$3*10^{-6}$	0.015	0.027
		Choplump	0.059	0.003	0.001	0.317
	$H_A^{(5)}$	Chang	$5*10^{-6}$	$2*10^{-13}$	$4*10^{-6}$	$7*10^{-6}$
		Choplump	$5.2*10^{-7}$	$3*10^{-19}$	$5*10^{-7}$	$5.3*10^{-7}$
	$H_A^{(6)}$	Chang	0.0006	$6*10^{-9}$	0.0004	0.001
		Choplump	0.004	$2*10^{-5}$	10^{-5}	0.041
3rd	H_0	Chang	0.496	0.042	0.015	0.772
		Choplump	0.510	0.039	0.004	0.800
	$H_A^{(1)}$	Chang	0.021	10^{-6}	0.016	0.026
		Choplump	$2*10^{-9}$	$7*10^{-24}$	$1.9*10^{-9}$	$2*10^{-9}$
	$H_A^{(2)}$	Chang	0.109	$7*10^{-5}$	0.087	0.142
		Choplump	0.210	0.028	0.003	0.965
	$H_A^{(3)}$	Chang	$2*10^{-5}$	$6*10^{-12}$	$2*10^{-5}$	$3*10^{-5}$
		Choplump	$6.5*10^{-10}$	$3*10^{-25}$	$6.4*10^{-10}$	$6.7*10^{-10}$
	$H_A^{(4)}$	Chang	0.003	10^{-7}	0.002	0.004
		Choplump	0.016	0.0003	$5*10^{-5}$	0.144
	$H_A^{(5)}$	Chang	$6*10^{-9}$	$6*10^{-19}$	$4*10^{-9}$	$9*10^{-9}$
		Choplump	$1.3*10^{-10}$	10^{-26}	$1.28*10^{-10}$	$1.31*10^{-10}$
	$H_A^{(6)}$	Chang	10^{-5}	$3*10^{-12}$	$7*10^{-6}$	$2*10^{-5}$
		Choplump	0.0002	$8*10^{-8}$	$2*10^{-7}$	0.003
4th	H_0	Chang	0.520	0.044	0.022	0.802
		Choplump	0.496	0.043	0.002	0.799
	$H_A^{(1)}$	Chang	0.001	$5*10^{-9}$	0.0007	0.0013
		Choplump	$1.09*10^{-17}$	$5*10^{-40}$	$1.07*10^{-17}$	$1.1*10^{-17}$
	$H_A^{(2)}$	Chang	0.019	$3*10^{-6}$	0.013	0.026
		Choplump	0.063	0.003	0.0003	0.375
	$H_A^{(3)}$	Chang	$2*10^{-9}$	10^{-19}	$2*10^{-9}$	$4*10^{-9}$
		Choplump	$1.41*10^{-18}$	$5*10^{-43}$	$1.4*10^{-18}$	$1.43*10^{-18}$
	$H_A^{(4)}$	Chang	$3*10^{-5}$	$2*10^{-11}$	$2*10^{-5}$	$5*10^{-5}$
		Choplump	0.001	10^{-6}	$3*10^{-7}$	0.0176

	$H_A^{(5)}$	Chang	$3*10^{-16}$	10^{-30}	$3*10^{-16}$	$4*10^{-16}$
		Choplump	$4.7*10^{-20}$	$7*10^{-46}$	$4.69*10^{-20}$	$4.74*10^{-20}$
	$H_A^{(6)}$	Chang	$3*10^{-10}$	$6*10^{-21}$	10^{-10}	$6*10^{-10}$
		Choplump	$7*10^{-8}$	$4*10^{-14}$	$3*10^{-12}$	$2.7*10^{-6}$
5th	H_0	Chang	0.499	0.038	0.019	0.786
		Choplump	0.491	0.042	0.001	0.799
	$H_A^{(1)}$	Chang	$2*10^{-6}$	$4*10^{-14}$	10^{-6}	$2*10^{-6}$
		Choplump	$6.3*10^{-34}$	$8*10^{-73}$	$6.2*10^{-34}$	$6.5*10^{-34}$
	$H_A^{(2)}$	Chang	0.001	10^{-8}	0.0005	0.0012
		Choplump	0.008	0.0002	$6*10^{-6}$	0.173
	$H_A^{(3)}$	Chang	0**	0	0	0
		Choplump	$1.08*10^{-35}$	$2*10^{-77}$	$1.07*10^{-35}$	$1.1*10^{-35}$
	$H_A^{(4)}$	Chang	$2*10^{-9}$	$3*10^{-19}$	10^{-9}	$5*10^{-9}$
		Choplump	$8*10^{-7}$	$7*10^{-12}$	$9*10^{-12}$	$4*10^{-5}$
	$H_A^{(5)}$	Chang	0	0	0	0
		Choplump	$1.12*10^{-38}$	$3*10^{-38}$	$1.11*10^{-38}$	$1.14*10^{-38}$
	$H_A^{(6)}$	Chang	0	0	0	0
		Choplump	$9*10^{-15}$	10^{-27}	$8*10^{-21}$	$5*10^{-13}$

(*):$H_0^{(1)}$:no risk reduction and no severity reduction , $H_A^{(1)}$:reduction rate 30% with severity reduction, $H_A^{(2)}$:reduction rate 30% without severity reduction, $H_A^{(3)}$: reduction rate 50% with severity reduction, $H_A^{(4)}$:reduction rate 30% without severity reduction, $H_A^{(5)}$:reduction rate 70% with severity reduction and $H_A^{(6)}$:reduction rate 70% without severity reduction.

(**): Not exactly zero, but a value very close to zero.



(5a) **(5b)**

Figure 5: 3-D Scatterplots for the mean p-values under each scenario

The figure that follows describes the distribution of the p-values under each hypothesis and each test statistic, with restriction to the first case for the sample size, i.e. N=1,000. Once again, the difference not only between the two groups, but also between the alternative hypotheses themselves cannot be questioned.

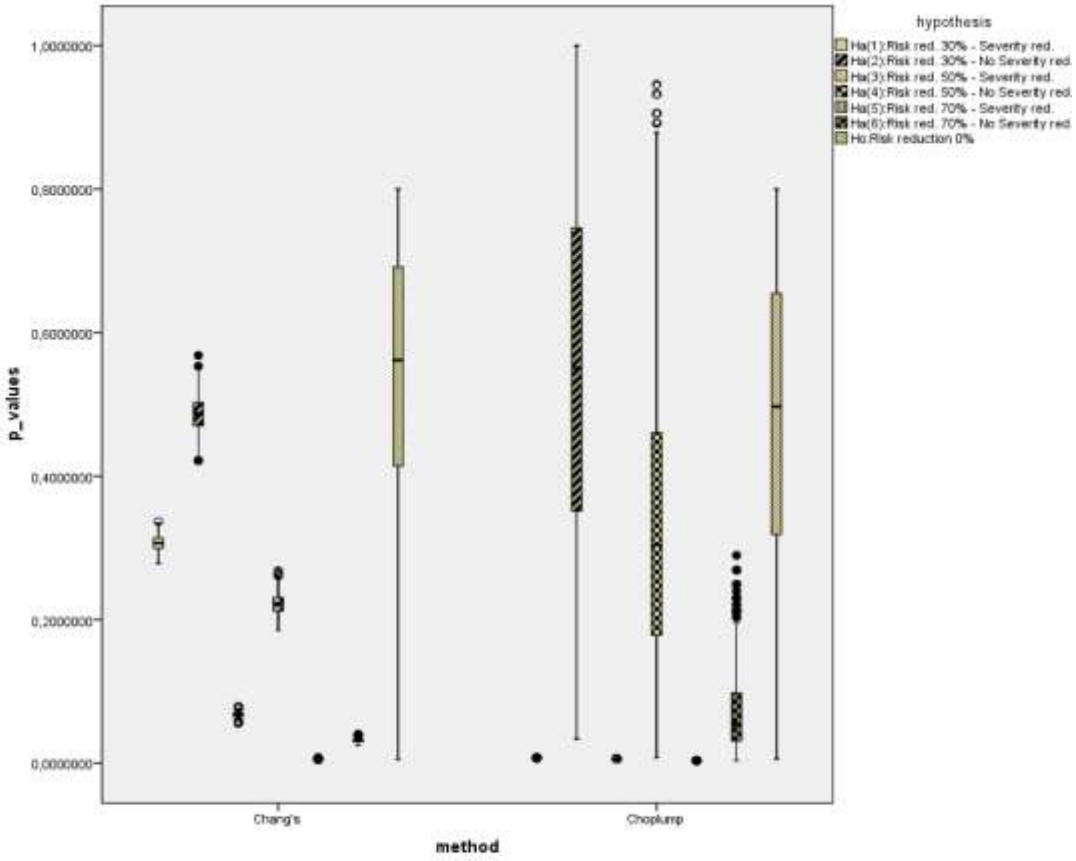


Figure 6: Boxplots of the p-values under each test statistic for N=1,000

The initial reason for carrying out all the above mentioned tests and comparisons was to estimate the two type error probabilities. After having obtained all the p-values, the type I error can be defined as the percentage of the cases for which the null hypothesis was rejected. Generally, this probability is preferred to be below 5%. Respectively, for the estimation of the type II error probability, the percentage of the cases for which the null hypothesis was not rejected should be considered.

In table 8, these probabilities are available for each different scenario. For both methods, the type I error remains at low levels, lower than 0.05, with the Choplump test yielding somewhat higher estimates. Even though, the scores used for the null hypothesis were from the same range, small differences ascribed to the simulation procedure, are detected more often as significant from the Choplump statistic. Concerning the type II error – and by extension the power - , what can be concluded from the table is that for small sample sizes the statistic of the Chang method fails to detect the underlying differences among the treated groups. On the other hand, the Choplump test seems to be able to find these differences even if the population's size is rather small.

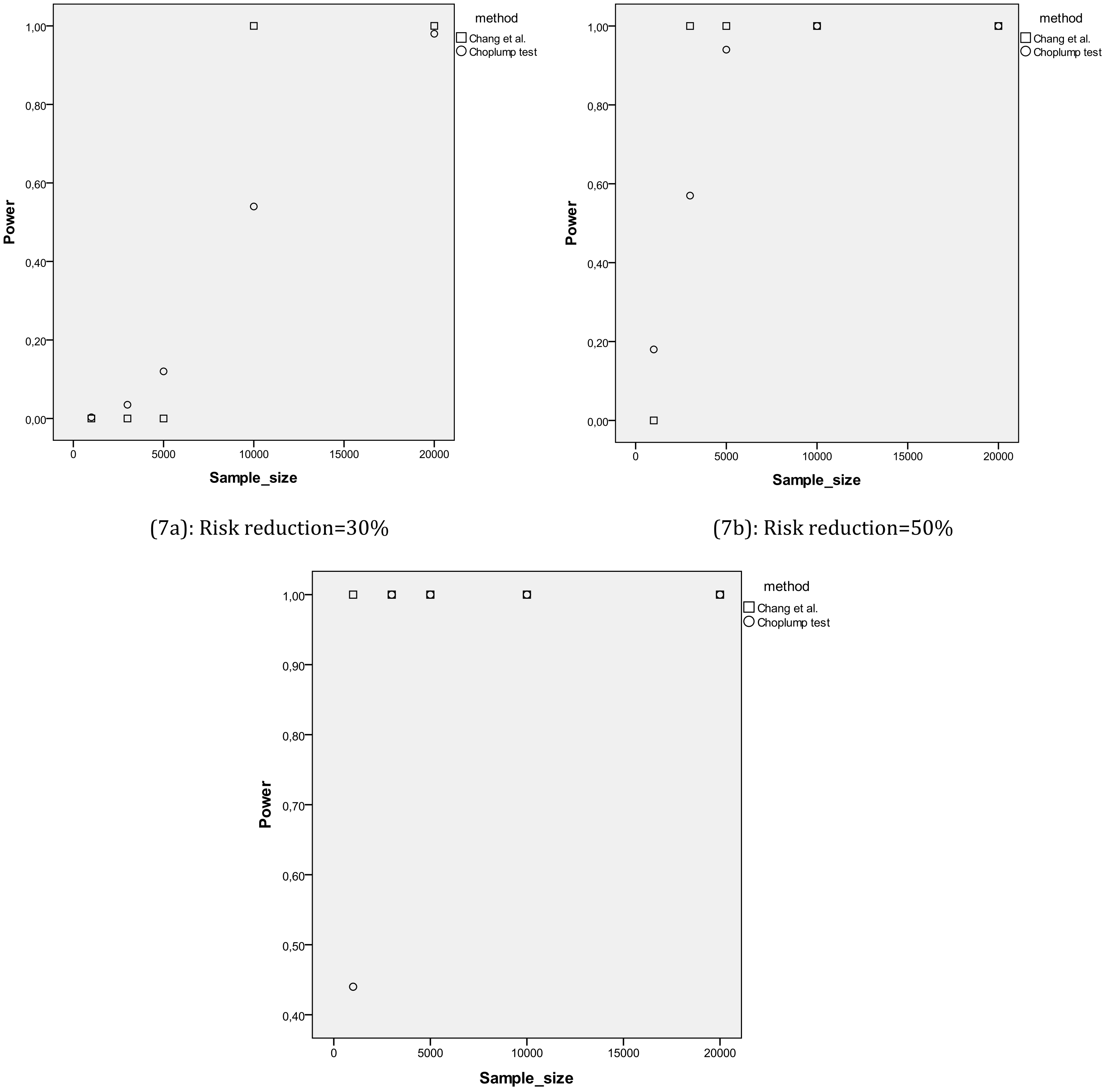
Table 8: Type I error, Type II error & Power of the tests

Scenario	RR*	SR**	Type I error		Type II error		Power	
			Chang	Choplump	Chang	Choplump	Chang	Choplump
1st	30%	Yes	0.01	0.02	1.00	0.00	0%	100%
2nd		No			1.00	0.997	0%	0.3%
3rd	50%	Yes			1.00	0.00	0%	100%
4th		No			1.00	0.82	0%	0.18%
5th	70%	Yes			0.00	0.00	100%	100%
6th		No			0.00	0.56	100%	44%
7th	30%	Yes	0.011	0.027	1.00	0.00	0%	100%
8th		No			1.00	0.965	0%	3.5%
9th	50%	Yes			0.00	0.00	100%	100%
10th		No			0.00	0.43	100%	57%
11th	70%	Yes			0.00	0.00	100%	100%
12th		No			0.00	0.00	100%	100%
13th	30%	Yes	0.013	0.025	0.00	0.00	100%	100%
14th		No			1.00	0.78	0%	12%
15th	50%	Yes			0.00	0.00	100%	100%
16th		No			0.00	0.06	100%	94%
17th	70%	Yes			0.00	0.00	100%	100%
18th		No			0.00	0.00	100%	100%
19th	30%	Yes	0.02	0.025	0.00	0.00	100%	100%
20th		No			0.00	0.46	100%	54%
21st	50%	Yes			0.00	0.00	100%	100%
22nd		No			0.00	0.00	100%	100%
23rd	70%	Yes			0.00	0.00	100%	100%
24th		No			0.00	0.00	100%	100%
25th	30%	Yes	0.026	0.032	0.00	0.00	100%	100%
26th		No			0.00	0.02	100%	98%
27th	50%	Yes			0.00	0.00	100%	100%
28th		No			0.00	0.00	100%	100%
29th	70%	Yes			0.00	0.00	100%	100%
30th		No			0.00	0.00	100%	100%

(*): RR stands for Risk Reduction, (**): SR stands for Severity Reduction

At this point, it is important to clarify the following. For the calculation of the tests, more important than the sample size (notated as N) is the number of cases in each group. But the number of cases in a group will always be much smaller compared to the number of individuals in that group (N/2) and this is due to the very small incidence rate (0.7%) that is taken into account. For this reason, the term N is used instead.

Concerning the three alternative hypotheses assuming that risk reduction comes in pair with severity reduction (i.e. $H_A^{(1)}$, $H_A^{(3)}$, $H_A^{(5)}$), there is a clear superiority of the Choplump test over the alternative one. Nevertheless, this cannot be stated for the remaining three alternatives, for which reduction in the severity scores is not accounted (i.e. $H_A^{(2)}$, $H_A^{(4)}$, $H_A^{(6)}$). For this reason, and since a graphical presentation is always more helpful, the scatterplots of the power of the two statistics under each case of risk reduction were plotted *(figure 7)*.



(7a): Risk reduction=30%

(7b): Risk reduction=50%

(7c): Risk reduction=70%

Figure 7: Scatterplots of power of the two tests under alternatives $H_A^{(2)}$, $H_A^{(4)}$ and $H_A^{(6)}$ where severity reduction is not considered to occur.

6. Discussion

In this paper, methods for dealing with datasets consisted of many zeros were presented and two of them were extendedly analyzed. In many cases the findings based on the different analyses and methods leaded to the conclusion that the Choplump test proposed by Follmann et al. (2009) is preferable over the Burden-of-Illness measure suggested by Chang et al. (1994) in some particular cases. Nevertheless, there is space left for some remarks to be pointed out.

Concerning the scores between the two groups under the alternative hypotheses, these might differ less, had larger ranges been used. For example, instead of the ranges 1-7 and 4-10 that were used for the vaccine and the placebo group in the illustration of section 5.2, respectively, the ranges 1-8 and 3-10 could have been considered as another approach. Thus, the overlapping of the distributions of the groups from the two groups would increase.

Judging from the statistics of the p-values (*table 7*), most interesting is the fact that the alternative hypotheses that also account for severity reduction, apart from risk reduction, present incredibly small distances between the minimum and the maximum values. This situation holds for both methods, but is seems to be more extreme in the case of the Choplump test. This finding can be graphically supported by figure 6, where for the Choplump test the boxplots can barely be visible. Conversely, the respective quantities are more far apart for the remaining three alternative hypotheses. On the outset, this phenomenon can be explained by the fact that in the first case the variability of the groups' scores is reduced compared to the relative variability of the competing set of hypothesis, because use of smaller ranges has been made. What is more, for an $H_A^{(i)}$ that accounts for severity reduction, the scores of the placebo and vaccine group will be far apart. On the other hand, for an $H_A^{(i)}$ not taking into account severity reduction, the scores will, on average, be close for the two groups. For a rough example to be given, from a randomly selected dataset – out of the simulated ones – and for $H_A^{(3)}$ (50% risk reduction and severity reduction) the mean score for the placebo group was 1274 and for the vaccine group it was 911, whereas for $H_A^{(4)}$ (50% risk reduction and no severity reduction) the relative means were 1003 and 998, respectively. Thus, differences are much more obvious, when severity reduction is assumed, and to this extension the p-values have a very small variability. Apart from that, as the sample size increases the decrease in the mean p-values is considerably higher and this phenomenon is especially noticeable with respect to the Choplump test. Last but not least, as the sample gets larger (N=20,000) and the risk reduction increases, Chang et al. method fails to calculate an exact value for the p-value and assigns the value zero.

In terms of power, it was proven that the Choplump test has a higher power even for a few incidents compared to Chang et al.'s test. Nevertheless, after increase of the sample size the second one manages to gain high power rather quickly. Also, the power of each test differs according to the type of the alternative hypothesis that is implied. Even the Choplump test, although it has a surprisingly high power, when severity reduction, additional to the risk reduction, is assumed, this fails to hold in the cases where this assumption is not considered. However, apart from that, a more interesting situation seems to hold when severity reduction is not assumed. To be more particular a higher power is observed for the Choplump test when the sample size is rather small (N=1,000) – although this is not always the case -, but as this increases (N=3,000-10,000) Chang's method increases rapidly in terms of power. Finally for very large samples sizes (N=20,000) both tests reach highest power. This phenomenon can be seen for the 6th, 10th, 16th, 20th and 26th scenarios, although for the last one the difference is very small and could be attributed to the simulation procedure. It looks as if there is range of moderate sample sizes for which the test of Chang et al. yields much higher power than the other one *(figure 7).* For the type I error probability, this was expected to be under 5% for most of the cases. The Choplump test yielded a higher estimation for this probability, due to its higher sensitivity in the sense that it considers as significant even very small deviations from the equality of the two groups' means.

As an overall conclusion, what can be stated is that both tests represent adequate approaches to the issue of handling a lot of zeros present in a study. However, the Choplump test can be generally characterized as dominant over its competitor only in cases when the efficacy of the vaccine is reflected by both risk and severity reduction, since as was proven earlier, when there is no reduction in the scores coming from the severity of the disease, the power yielded is, in some cases substantially, higher for the method proposed by Chang et al. Thus, it might be interesting for an analysis, based on more simulations, to be carried out in order to investigate this phenomenon more thoroughly.

Most interesting, of course, would be the application of the tests on real data when these are available. In that case the results would be much more reliable, since bias coming from the simulation procedure would not be present. Furthermore, the severity scores coming from a real trial might not follow the general philosophy that was adopted for the generation of the scores for the purposes of the analysis and different situations could arise.

References

Brazier J, Jones N, Kind P. **Testing the Validity of the EuroQoL and comparing it with the SF-36 health survey questionnaire**. Qual Life Res. 1993; 2(3):169-180

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.

Coplan PM, Schmader K, Nikas A, Chan IS, Choo P, Levin MJ, Johnson G, Bauer M, Williams HM, Kaplan KM, Guess HA, Oxman MN: **Development of a measure of the burden of pain due to herpes zoster and postherpetic neuralgia for prevention trials: adaptation of the brief pain inventory**. The Journal of Pain, Vol.5, No6 (August), 2004, pp. 344-356.

Follmann D.A., Fay M, & Proschan M. **"Chop-lump tests for vaccine trials"**, Biometrics 2009; 65 885-893.

Lachenbruch PA: **Analysis of data with clumping at zero.** *Biometrische Zeitschrift* 1976, 18:351-356.

Lachenbruch PA: **Comparison of two-part models with competitors.** *Statistics in Medicine* 2001, 20**:**1215-1234.

Lachenbruch PA: **Analysis of data with excess zeros.** *Statistical Methods in Medical Research* 2002, 11:297-302.

McDowell I, Newell C: Measuring Health: **A Guide to Rating Scales and Questionnaires**. New York, NY, Oxford University Press, 1996, p.350

Neuhäuser M, Boes T, Jöckel K-H: **Two-part permutation tests for DNA methylation and microarray data**. BMC Bioinformatics 6: 35 (2005)

Taylor S, Pollard K, **Hypothesis Tests for Point-Mass Mixture Data with Application to Omics Data with Many Zero Values**, Vol.8-Issue 1, 2009, Article 8.

Appendix

SAS codes – SAS macros

```
/*** HISTOGRAMS ***/
proc sort data=thesis.data;
by Z;
run;

proc univariate data=thesis.data noprint;
by Z;
*where W>0;
var W;
histogram / normal;
run;

proc means data=thesis.data mean std median mode var min max;
by Z;
*where W>0;
var W;
run;

/*** AREA UNDER THE CURVE ***/
proc gplot data=thesis.means;
plot mean_daily_BOI*day=Z/overlay;
symbol1 interpol=join color=black;
symbol2 interpol=join l=3 color=black;
run;

/************************************************************************/

Program: choplumpExact.sas

Author: Marie N. Kassapian

Date: August 26, 2011


This program calculates the exact 2-sided p-value of a choplump test for:
Ho: equality of the BOI scores between placebo and vaccine group

The main idea is based on the R code created by Michael P. Fay.


Input: A dataset containing two columns.A column indicating the BOI scores
of the patients and a column indicating the group membership of them.

Output: A 2-sided p-value for testing the null hypothesis of the equality
of the BOI scores among the treated groups.

Temporary Datasets: Dout, Rank, Nties, Allperm, Output, Trans, New,

This programs calls the following macros:

  i. %wilcoxon(d,Z)

 ii. %comb(t,n)

iii. %chopgeneral(d,W,Z)
```

```
 iv. %choplumpExact(d,W,Z)
/********************************************************************/
/******** CALCULATE THE TEST STATISTIC OF THE OBSERVED DATA ********/
/********************************************************************/

%macro wilcoxon(d,Z);

  proc rank data=&d out=Ranks ;
    var W;
    ranks Ranks;
  run;
  proc sql noprint;
    select count(*) into: N1
      from &d
      where &Z = 1;
      select count(*) into: N0
      from &d
      where &Z = 0;
  quit;

  proc sql noprint;
    select (sum(Ranks)- &N0 * (&N0 + 1)/2) into: Statistic
      from Rank
    where &Z=0;
  run;

  proc freq data=Rank noprint;
    tables Ranks/out=Nties;
  run;

  proc sql noprint;
   select (sqrt((&N0*&N1/12)*((&N0+&N1+1)-sum(COUNT*COUNT*COUNT -COUNT)/
          ((&N0+&N1)*(&N0+&N1-1))))) into: Sigma
      from Nties;
  quit;

  %global Out;
  data _NULL_;
    out=(&Statistic - &N0 * &N1/2)/&Sigma;
    call symput('Out',Out);
  run;

%mend wilcoxon;

/*************************************************************************/
/**** CREATE A MATRIX WITH ALL POSSIBLE PERMUTATIONS BASED ON t TOTAL ****/
/******************* PATIENTS & n TOTAL CONFIRMED CASES ******************/
/*************************************************************************/

 %macro comb(t,n);

  %local i n;
  data allperm;
  %do i=1 %to &t.;
  do i&i.=0 to 1;
  %end;
  if sum(of i1-i&t.)=&n. then output;
  %do i=1 %to &t.;
  end;
  %end;
  run;
  data Allperm;
   set Allperm;
      n=_N_; run;

%mend comb;

/***********************************************************************/
/********************** CREATE THE CHOPPED DATASET *********************/
/***********************************************************************/

%macro chopgeneral(d,W,Z);

  proc sort data=&d;
    by &W &Z;
  run;
  proc sql noprint;
    select count(*) into: M
      from &d
      where &W ne 0;
    select count(*) into: M0
      from &d
      where &W ne 0
      and &Z = 0;
    select count(*) into: M1
      from &d
      where &W ne 0
      and &Z = 1;
    select count(*) into: N1
      from &d
      where &Z = 1;
    select count(*) into: N
      from &d;
      select count(*) into: N0
      from &d
      where &Z = 0;
  quit;

  data _NULL_;
   k0 = %eval(&N0 - &M0);
   k1 = %eval(&N1 - &M1);
      call symput('k0',k0);
      call symput('k1',k1);
  run;

  /* do the chopping */
  %if &k0 > &k1 %then %do;
    data dout(drop=t);
        set &d;
        if &Z=1 and &W=0 then delete;
        retain t 0;
        if &Z=0 then do;
          if &k1 ne t then do;
          t+1;
              delete;
          end;
        end;
      run;
  %end;
  %else %if &k0 < &k1 %then %do;
    data dout(drop=t);
```

```
        set &d;
        if &Z=0 and &W=0 then delete;
        retain t 0;
        if &Z=1 then do;
          if &k0 ne t then do;
          t+1;
              delete;
          end;
        end;
      run;
  %end;
  %else %do;
    data dout;
        set &d(where=(&W ne 0));
      run;
  %end;

%mend chopgeneral;

/***********************************************************************/
/********************** CALCULATE THE EXACT P-VALUE ********************/
/***********************************************************************/

%macro choplumpExact(d,W,Z);

  proc sort data=&d;
    by &W &Z;
  run;

  %chopgeneral(&d,W,Z);               /* output dataset is called dout */
  %wilcoxon(dout,Z);                  /* output variable is called out */

  data _NULL_;
   T0= &out;
      call symput('T0',T0);
  run;

  proc sql noprint;
    select count(*) into: M
      from &d
      where &w ne 0;
    select count(*) into: M0
      from &d
      where &w ne 0
      and &z = 0;
    select count(*) into: M1
      from &d
      where &w ne 0
      and &z = 1;
    select count(*) into: N1
      from &d
      where &z = 1;
    select count(*) into: N
      from &d;
      select count(*) into: N0
      from &d
      where &z = 0;
  quit;

  data _NULL_;
   k0 = %eval(&N0 - &M0);
   k1 = %eval(&N1 - &M1);
    call symput('K0',k0);
    call symput('K1',k1);
  run;

 %comb(&N,&N1);                    /* output dataset is called Allperm */

  proc sql noprint;
    select count(*) into: Nperm
      from Allperm;
  quit;
  %put &Nperm;

  %do i=1 %to &Nperm;
  data Output;
   set Allperm(where=(n=&i));
  run;

  proc transpose data=output out=Trans;
   var _all_;
  run;

  data Trans;
   set Trans;
    if _name_='n' then delete;
  run;
  data New;
   merge &d Trans;
  run;

  %chopgeneral(New,W,COL1);           /* output dataset is called Dout */
  %wilcoxon(Dout,COL1);               /* output variable is called Out */

  data _NULL_;
   T&i= &Out;
    call symput("T&i",T&i);
  run;
  %end;

  data _NULL_;
   plower=0;
   pupper=0;
    %do i=1 %to &nperm;
      if &&T&i LE &T0 then plower=plower+1;
      if &&T&i GE &T0 then pupper=pupper+1;
    %end;
      call symput('plower',plower);
      call symput('pupper',pupper);
  run;
  data _NULL_;
   p_lower =&plower./&Nperm.;
   p_upper =&pupper./&Nperm.;
   p_2sided =min(1, 2 * min(&plower.,&pupper.)/&Nperm.);
      call symput('p_lower',p_lower);
      call symput('p_upper',p_upper);
      call symput('p_2sided',p_2sided);
  run;
  %put p_lower=&p_lower;
  %put p_upper=&p_upper;
  %put p_2sided=&p_2sided;
```

```
%mend choplumpExact;
/************************************************************************/

Program: choplumpApprox.sas

Author: Marie N. Kassapian

Date: August 26, 2011


This program calculates the approximated 2-sided p-value of a choplump test
for the null hypothesis:
Ho: equality of the BOI scores between placebo and vaccine group


The main idea is based on the R code created by Michael P. Fay.


Input: A dataset containing two columns. A column indicating the BOI scores
of the patients and a column indicating the group membership of them.

Output: A 2-sided p-value for testing the null hypothesis of the equality
of the BOI scores among the treated groups.

Temporary Datasets: Repeat1, Reapeat2, A, Repeat3, Repeat4, B, Sort, D2,
D3, Rename, R1, SM, D2, D3, Dout, Dchop, H, Ti_&hfirst : Ti_&hlast, Final


This programs calls the following macros:

  i. %chopgeneral(d,W,Z)

 ii. %TDiM(d,W,Z)

iii. %Qh(d,W,Z,h,N1,N0,M,SM,T0,use_ranks)

 iv. %mergedata

  v. %choplumpApprox(d,W,Z,use_ranks= )


/*************************************************************************/
/*********************** CREATE THE CHOPPED DATASET **********************/
/*************************************************************************/

%macro chopgeneral(d,W,Z);
  proc sort data=&d;
    by &W &Z;
  run;
  proc sql noprint;
    select count(*) into: M
      from &d
      where &W ne 0;
    select count(*) into: M0
      from &d
      where &W ne 0
      and &Z = 0;
    select count(*) into: M1
      from &d
      where &W ne 0
      and &Z = 1;
    select count(*) into: N1
      from &d
      where &Z = 1;
    select count(*) into: N
      from &d;
      select count(*) into: N0
      from &d
      where &Z = 0;
  quit;

  data _NULL_;
    k0 = %eval(&N0 - &M0);
      k1 = %eval(&N1 - &M1);
       call symput('K0',k0);
       call symput('K1',k1);
  run;

  /* do the chopping */
  %if &k0 > &k1 %then %do;
    data dout(drop=t);
        set &d;
        if &Z=1 and &W=0 then delete;
        retain t 0;
        if &Z=0 then do;
          if &k1 ne t then do;
          t+1;
              delete;
          end;
        end;
      run;
  %end;
  %else %if &k0 < &k1 %then %do;
    data dout(drop=t);
        set &d;
        if &Z=0 and &W=0 then delete;
        retain t 0;
        if &Z=1 then do;
          if &k0 ne t then do;
          t+1;
              delete;
          end;
        end;
      run;
  %end;
  %else %do;
    data dout;
        set &d(where=(&W ne 0));
      run;
  %end;
%mend chopgeneral;

/***********************************************************************/
/********** ILLUSTRATION OF PERMUTATIONAL CENTRAL LIMIT THEOREM ********/
/***********************************************************************/

%macro TDiM(d,W,Z);

 proc sql noprint;
    select count(*) into: L
      from &d;
      select sum(&W*&Z) into: sumWZ
      from &d;
      select mean(&W) into: meanW
      from &d;
    select mean(&Z) into: meanZ
      from &d;
      select std(&W) into: stdW
      from &d;
      select std(&Z) into: stdZ
      from &d;
 quit;

 data _NULL_;
  T0=(1/sqrt(&L-1))*(&sumWZ - &L*&meanW*&meanZ)/(&stdW*&stdZ);
   call symput("T0",T0);
 run;
 %put T0=&T0;

%mend TDiM;

/************************************************************************/
/************* CALCULATE THE Qh-HAT FOR A SPECIFIC VALUE OF H ***********/
/************************************************************************/

%macro Qh(d,W,Z,h,N1,N0,M,SM,T0,use_ranks);

proc sql noprint;
 select count(*) into: N
 from &d;
quit;

    %let K = %eval(&N1+&N0-&M);
      %let k1p = %eval(&h);
      %let k0p = %eval(&K-&h);
      %let M1p = %eval(&N1-&k1p);
      %let M0p = %eval(&N0-&k0p);

/* do the chopping */
  data _NULL_;
   if &M0p/&N0 >= &M1p/&N1 then do;
     k0c=0 ;
     k1c=&k1p-floor(&N1*&k0p/&N0);
       end;
   else if &M0p/&N0 < &M1p/&N1 then do;
     k0c=&k0p-floor(&N0*&k1p/&N1);
     k1c=0;
   end;
    call symput('k0c',k0c);
    call symput('k1c',k1c);
   run;


  %let Kstar = %eval(&k0c+&k1c);
  %let M0c = %eval(&M0p);
  %let M1star = %eval(&M1p);
  %let hstar = %eval(&k1c);
  %let M1star = %eval(&N1-&h);
  %let N1star = %eval(&M1star+&hstar);
  %let N0star = %eval(&M0c+&k0c);

 %if &use_ranks=1 %then %let S0Kstar=%sysevalf(-(%sysevalf(&Kstar)-1)/2);
  %else %if &use_ranks ne 1 %then %let S0Kstar=0;
 /* create dataset A consisting of variable ‘a’ with elements:
    rep(1,n1star),rep(0,n0star) */
 data repeat1;
  keep a;
  do i=1 to &N1star;
  a=1;output;
 end;
 data repeat2;
  keep a;
  do i=1 to &N0star;
  a=0;output;
 end;

 data A;
   set repeat1 repeat2;
 run;

/* create dataset B consisting of variable ‘b’ with elements:
   rep(S0Kstar,Kstar),SM */
 data repeat3;
  keep b;
  do i=1 to &Kstar;
  b=&S0Kstar;output;
 end;
 data repeat4;
  set SM;
  keep b;
  b=SM;output;
 run;

 data B;
   set repeat3 repeat4;
 run;

proc sql noprint;
 select var(a) into: vara
  from A;
 select var(b) into: varb
  from B;
 select sum(SM) into: sumSM
  from SM;
 select var(SM) into: varSM
  from SM;
 select mean(SM) into: meanSM
  from SM;
quit;

%let Nstar = %eval(&M+&Kstar);
%let SbarNstar = %sysevalf((&Kstar*&S0Kstar+&sumSM)/&Nstar);
%let VM = %sysevalf((&M-1)*&varSM*&vara);
%let VNstar = %sysevalf((&Nstar-1)*&varb*(&Nstar-
&N1star)*&N1star/(&Nstar*(&Nstar-1)));

data _NULL_;
 Zstat = ((&T0*sqrt(&VNstar)-&hstar*&S0Kstar+&N1star*&SbarNstar -(&N1star-
          &hstar)*&meanSM)/sqrt(&VM));
 Qhhat = (probnorm(Zstat));
 Dhyper = (PDF('HYPER',&h,&N,&K,&N1));
  call symput('Zstat',Zstat);
  call symput('Qhhat',Qhhat);
  call symput ('Dhyper',Dhyper);
run;

/* calculate the probability of a permutation having h zeros in the vaccine
   group multiplied with the Qhhat (for a specific value of h) */
data Ti_&h;
 Ti=&Qhhat*&Dhyper;
 call symput('Ti',Ti);
run;

%mend Qh;

/**************************** MERGE DATASETS **************************/

%macro mergedata;
 %do i=&hfirst %to &hlast;
  proc append base=Final data=Ti_&i;
  run;
 %end;

%mend mergedata;

/***********************************************************************/
/**************** CALCULATE THE APPROXIMATED P-VALUE ******************/
/***********************************************************************/

%macro choplumpApprox(d,W,Z,use_ranks=1);    /***  parameter 'use_ranks'
takes value 1 if ranking is used or 0 if ranking is not used  ***/

  proc sort data=&d out=Sort;
    by &W &Z;
  run;
  proc sql noprint;
    select count(*) into: M
      from &d
      where &W ne 0;
    select count(*) into: M0
      from &d
      where &W ne 0
      and &Z = 0;
    select count(*) into: M1
      from &d
      where &W ne 0
      and &Z = 1;
    select count(*) into: N
      from &d;
    select count(*) into: N1
      from &d
      where &Z = 1;
    select count(*) into: N0
      from &d
      where &Z = 0;
  quit;

   %let K = %eval(&N-&M);
   %let K0 = %eval(&N0-&M0);
   %let K1 = %eval(&N1-&M1);

 data D2;
  set &d;
  obs=_n_;
 run;
proc sql noprint;
 create table D3 as
 select *
 from D2
 where (&N-&M+1)<=obs<=(&N);   /*select elements in positions N+M-1 to N */
 run;

data Rename;
set D3;
rename W=WM Z=ZM;
run;

proc rank data=Rename out=R1 ;
    var WM;
    ranks RankWM;
  run;

data SM;
 set R1;
 keep SM;
 if &use_ranks=1 then SM = RankWM;
 else if &use_ranks ne 1 then SM = WM;
 call symput('SM',SM);
run;

%chopgeneral(&d,W,Z);                  /* output dataset is called dout */

 proc rank data=Dout out=Dchop ;
  var W;
  ranks RankW; run;

/* calculate the test statistic for the observed data */
%global T0;
%if &use_ranks=1 %then %TDiM(Dchop,RankW,Z);
%else %if &use_ranks ne 1 %then %TDiM(Dchop,W,Z);
%put &T0;

/* create a variable with all possible values for h (h=number of zeros in
   the vaccine group) */
data _NULL_;
 hfirst = max(0,&N1-&M);
 hlast = min(&N1,&K);
  call symput('hfirst',hfirst);
  call symput('hlast',hlast);
run;

data H;
  keep h;
  do i=&hfirst to &hlast;
  h=i;output;
 end;
run;

/*** apply macro %Qh on all possible values of h ***/
filename TMP_FIL TEMP;
data _NULL_;
set H;
file TMP_FIL;
put '%Qh(&d,W,Z,' h ',&N1,&N0,&M,SM,&T0,&use_ranks);' ;
run;
```

```
%include TMP_FIL;
%mergedata;

proc sql noprint;
 select sum(Ti) into:sum
  from Final;
quit;

data _NULL_;
 p_lower = (1-&sum);
 p_upper = (&sum);
 min=min(p_lower,p_upper);
  call symput('p_lower',p_lower);
  call symput('p_upper',p_upper);
  call symput('min',min);
run;
data _NULL_;
 p_2sided=min(1,2*&min) ;
  call symput('p_2sided',p_2sided);
run;
  %put p_lower=&p_lower;
  %put p_upper=&p_upper;
  %put p_2sided=&p_2sided;

%mend choplumpApprox;


/**************************************************************************/

Program: chang.sas

Author: Marie N. Kassapian

Date: August 26, 2011


This program calculates the 2-sided p-value of the test suggested by Chang
et al. for the null hypothesis:
Ho: equality of the BOI scores between placebo and vaccine group

Input: A dataset containing two columns. A column indicating the BOI scores
of the patients and a column indicating the group membership of them.

Output: A 2-sided p-value for testing the null hypothesis of the equality
of the BOI scores among the treated groups.

Temporary Datasets: W0, W1, Score0, Score1

/**************************************************************************/

%macro Chang(d,W,Z);
 proc sql noprint;
    select count(*) into: M
      from &d
      where &w ne 0;
    select count(*) into: M0
      from &d
      where &w ne 0
      and &z = 0;
    select count(*) into: M1
      from &d
```

```
     where &w ne 0
     and &z = 1;
   select count(*) into: N
     from &d;
   select count(*) into: N1
     from &d
     where &z = 1;
   select count(*) into: N0
     from &d
     where &z = 0;
   select sum(&W) into: sumW0
   from &d
   where &W ne 0
    and &Z = 0;
   select sum(&W) into: sumW1
   from &d
   where &W ne 0
    and &Z = 1;
   select mean(&W) into: meanW0
   from &d
   where &W ne 0
    and &Z = 0;
   select mean(&W) into: meanW1
   from &d
   where &W ne 0
    and &Z = 1;
 quit;

 %let Tstat = %sysevalf((1/&N0)*&sumW0-(1/&N1)*&sumW1);

data W0;
 set &d;
 keep W;
 where W ne 0 & Z=0;
 rename W=W0;
  call symput('W0',W0);
run;

data W1;
 set &d;
 keep W;
 where W ne 0 & Z=1;
 rename W=W1;
  call symput('W1',W1);
run;

 %let p = %sysevalf(&M/&N);
 %let xbar = %sysevalf((&sumW0+&sumW0)/&M);

 data score0 ;
  set W0;
   score0 =(W0-&meanW0)**2 ;
    call symput('score0',score0);
 run;

 data score1 ;
  set W1;
   score1 =(W1-&meanW1)**2 ;
      call symput('score1',score1);
 run;
```

```
  proc sql noprint;
   select sum(score0) into: sumscore0
   from score0;
   select sum(score1) into: sumscore1
   from score1;
  quit;

  %let s0est=%sysevalf(&sumscore0/%eval(&M0-1));
  %let s1est=%sysevalf(&sumscore1/%eval(&M1-1));
  %let varThat = %sysevalf((&xbar**2)*&p*(1-&p)*(1/&N0+1/&N1)
                                              +&p*(&s0est/&N0+&s1est/&N1));

 data _NULL_;
  finStat = abs(&Tstat/sqrt(&varThat));
   call symput('finStat',finStat);
 run;

 data _NULL_;
  p2sided = 2*(1-probnorm(&finstat));
   call symput('p2sided',p2sided);
 run;

%mend Chang;

/************************************************************************/
/***************** IMPLEMENTATION OF CHI-SQUARE TEST FOR ****************/
/*********** Ho: EQUALITY OF INCIDENT RATES BETWEEN THE GROUPS **********/
/************************************************************************/

proc sql ;
create table Table1 as
 select *
 from thesis.datanew
 where W ne 0
  and Z=0;
quit;

proc sql ;
create table Table2 as
 select *
 from thesis.datanew
 where W ne 0
  and Z=1;
quit;

data Table3;
 set Table1 Table2;
run;

proc freq data=Table3;
tables Z / chisq; run;

/************************************************************************/
/********************** IMPLEMENTATION OF T-TEST FOR ********************/
/************* Ho: EQUALITY OF SEVERITY SCORES BETWEEN THE GROUPS *******/
/************************************************************************/

proc ttest data=Table3;
 class Z;
 var W;run;
```

```
/*************************************************************************/
/*** R CODE FOR THE IMPLEMENTATION OF CHANG ET AL. TEST STATISTIC FOR ***/
/************ Ho:EQUALITY OF B.O.I.SCORES BETWEEN THE GROUPS ************/
/*************************************************************************/

chang<-function(W,Z){

N0<-length(Z[Z==0])
N1<-length(Z[Z==1])
N<-N0+N1

M<-length(W[W!=0])
M0<-length(W[W!=0 & Z==0])
M1<-length(W[W!=0 & Z==1])

p<-M/N
W0<-W[W!=0 & Z==0]
W1<-W[W!=0 & Z==1]

sumW0<-sum(W0)
sumW1<-sum(W1)

xbar<-((sumW0+sumW1)/M)

s02<-(sum((W0-mean(W0))^2))/(M0-1)
s12<-(sum((W1-mean(W1))^2))/(M1-1)

T<-(1/N0)*sumW0-(1/N1)*sumW1
VT<-((xbar^2)*p*(1-p)*(1/N0+1/N1)+p*(s02/N0+s12/N1))
test<-abs((T/sqrt(VT)))

p2sided<- 2*(1-pnorm(test))
p2sided
}


/*************************************************************************/
/************** R CODE FOR THE SIMULATION OF THE DATASETS ***************/
/*************************************************************************/

N<-
a<-
b<-

Z<-c(rep(0,N-a),rep(1,N-b),rep(0,a),rep(1,b))    #fix membership indicator
data1<-matrix(numeric(0),a,182)
for (i in 1:182) {
data1[,i]<-sample(c(1:10),a,replace=T) #sample the scores for placebo group
}
data2<-matrix(numeric(0),b,182)
for (i in 1:182) {
data2[,i]<-sample(c(1:7),b,replace=T)  #sample the scores for vaccine group
}
data3<-matrix(rep(0,(2*N-a-b)*182),2*N-a-b,182)
score<-rbind(data3,data1,data2)                          #merge all scores
W<-apply(score,1,sum)                  #compute total score for each patient
finaldata<-cbind(W,Z)                                          #final dataset
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Analysis of Methods Used in Trials Containing Many Zeros

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt
behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -,
vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten
verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Kassapian, Mari

Datum: **15/09/2011**