A STUDY OF MULTI-GROUP AND MULTIVARIATE TESTS OF EQUIVALENCE WITH APPLICATION TO MICROARRAY AND CLINICAL TRIAL DATA

by

CELESTE YANG

XIANGQIN CUI, CHAIR
AL BARTOLOCCI, CO-CHAIR
IMMACULADA (CHI-CHI) ABAN
DEGUI ZHI
LISA SCHWIEBERT
VICTOR DARLEY-USMAR

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

BIRMINGHAM, ALABAMA

2013
A STUDY OF MULTI-GROUP AND MULTIVARIATE TESTS OF EQUIVALENCE WITH APPLICATION TO MICROARRAY AND CLINICAL TRIAL DATA

CELESTE YANG

PHD IN BIOSTATISTICS ABSTRACT

In this dissertation we have adapted two multi-group equivalence tests to be performed on high-dimensional data. The F and Range test for multi-group equivalence (Wellek 2010) were applied to the public microarray dataset GSE11291 (Barger et al. 2008) to detect equivalently expressed genes. They were also evaluated in terms of type I error and power using single gene simulation and a high-dimensional simulation. The F test has higher power than the Range for the same simulated data and parameter settings, in the single gene simulations. The power of the two tests is similar in the high-dimension simulation.

The multi-group equivalence tests that were applied to the microarray dataset were used to create the R package \textit{EquivMulti}. The package offers multi-group testing utilizing the F and Range tests of equivalence with the adaptations discussed in

Finally, in a clinical trial setting we use simulation study and real data analysis to analyze the behavior of two existing multivariate equivalence tests, Hotellings $T^2$ test and the intersection union test (IUT). We also propose an alternative multivariate equivalence test, Max Test, which we evaluate along with the others using simulation study and real data analysis.
ACKNOWLEDGEMENTS

I would just like to thank all the people who helped me during and through the long process of writing this dissertation, in particular I am grateful to my Chair, Dr. Cui, and Co-Chair Dr. Bartolucci for being willing to work with me long distance and in unusual circumstances. I am also grateful to my committee members Drs. Chi-Chi Aban, Lisa Schwiebert, Degui Zhi, and Dr. Darley-Usmar for their patience and helpful critiques at my proposal and my progress presentation. The entire Biostatistics department, and above all Drs. Aban, Cutter, and Tiwari have been supportive of me during the hardest time in my life, the loss of my dad to metastatic prostate cancer. It’s been a long struggle for me to just write out the work I did, and I words are not enough for the debt of thanks I owe to my mentors Dr. Cui and Dr. Bartolucci for their patience with me. Additionally, I am very grateful to Dr. Aban’s firm, but kind words that helped me pick myself up and keep going. This dissertation has been, and still is, a work in progress, and a work in progress in my own self.

THROW YOURSELF LIKE SEED
- Miguel de Unamuno, translated by Robert Bly

Shake off this sadness, and recover your spirit;
sluggish you will never see the wheel of fate
that brushes your heel as it turns going by,
the man who wants to live is the man in whom life is abundant.

Now you are only giving food to that final pain
which is slowly winding you in the nets of death,
but to live is to work, and the only thing which lasts
is the work; start then, turn to the work.

Throw yourself like seed as you walk, and into your own field,
don’t turn your face for that would be to turn it to death,
and do not let the past weigh down your motion.

Leave what’s alive in the furrow, what’s dead in yourself,
for life does not move in the same way as a group of clouds;
from your work you will be able one day to gather yourself.
TABLE OF CONTENTS

ABSTRACT .................................................................................................................................................. ii

ACKNOWLEDGEMENTS ............................................................................................................................... iii

LIST OF TABLES ............................................................................................................................................... iv

LIST OF FIGURES ........................................................................................................................................ vi

INTRODUCTION ............................................................................................................................................... 1

MULTI-GROUP EQUIVALENCE TESTS AND THEIR APPLICATION TO MICROARRAY DATA .......................................................... 16

EQUIVMULTI: A MULTI-GROUP EQUIVALENCE TEST PACKAGE ................................................................. 61

MULTIVARIATE EQUIVALENCE TESTS WITH APPLICATION TO CLINICAL TRIAL DATA .......................................................... 68

SUMMARY .................................................................................................................................................... 101

DISCUSSION ............................................................................................................................................... 106

GENERAL LIST OF REFERENCES ................................................................................................................ 111
LIST OF TABLES

MULTI-GROUP EQUIVALENCE TESTS AND THEIR APPLICATION TO MICROARRAY DATA

1 Power of the F test as a function of the difference ratio (DR) and simulated mean value. Group size is set as 4; variance is set as 0.12; and sample size is set as 6…………………………………………………………………………………………………35

2 Type I error for the F and Range tests from Single-Gene Simulations. k, denotes number of groups; and n, denotes sample size. Difference ratio (DR) is set for 1.55. Type I error rate is set as 0.05……………………………………………………………………………37

3 Power of the range test for as a function of the difference ratio (DR) and the simulated means when the group size, k=4, and variance=0.12………………………………………43

4 Power of the Range test is affected by Group Numbers. The setting of the other parameters are: difference ratio, 1.55; variance, 0.12; and sample size, 6. SMax is the effect size of the range test given in equation 6. The mean value is the value assigned to the group means in the single gene simulations………………………………………44

4 Equivalently Expressed Genes identified by the F and Range Test. The DR was varied from 1.25 to 1.85. The adjusted alpha was set at 0.05. The counts are given before and after the Benjamini and Hochberg adjustment………………………………………49

EQUIVMULTI: A MULTI-GROUP EQUIVALENCE TEST PACKAGE

1 The number of equivalent genes before and after FDR adjustment using the Benjamini and Hochberg method. DR is the difference ratio criterion given in paper 1……………………………………………………………………………………………66
SUMMARY OF MULTIVARIATE EQUIVALENCE TESTS INVESTIGATED IN THIS PAPER

1. Type I errors estimated from the simulation data. “Equal”, “Real”, and “One” refer to the mean pattern generation described in Simulation section 1. Alpha is set to 0.05. The equivalence limit for the data generation was set to be 40 for the Max Test and the IUT and varied for the Hotellings T2 so that the equivalence limit would be as close to the estimated MHD as possible for each unique parameter setting.
LIST OF FIGURES

MULTI-GROUP EQUIVALENCE TESTS AND THEIR APPLICATION TO MICROARRAY DATA

1 Power curves for the F test from single gene simulations. The columns represent the different group sizes (k=3, 4, 5), and the rows represent different variances ($\sigma^2=0.04, 0.12, 0.24$). The sample sizes (N values) are indicated by the legend in the first panel. The simulated mean value represents the group mean values denoted as “a” throughout the text. The difference ratio (DR) is set as 1.55 .................................................................36

2 The Power of the F test is affected by number of groups. In this figure the power is given for a difference ratio criterion of 1.55, variance of 0.12, and sample size of 6. k, group numbers. Adjusted effect size is calculated as the square root of the noncentrality parameter given in equation (14).................................................................................................................39

3 Power for the Range test from the single-gene simulations. The columns represent the different group size, k=3, 4, 5, and the rows represent different variances $\sigma^2=0.04, 0.12, 0.24$. The sample sizes (N values) are indicated by the legend in the first panel. The simulated mean value represents the group mean values denoted as “a” throughout the text. The difference ratio (DR) is set as 1.55..................................................................................................................41

4 Power for the Range and F test for the High Dimension Gene Simulations. Colored lines represent different values of the difference ratio criterion. The “a” values are the simulated mean values........................................................................................................46
MULTIVARIATE EQUIVALENCE TESTS WITH APPLICATION TO CLINICAL TRIAL DATA

1 Mean P/F Ratio Values Over Time by Treatment Group. NO represents the inhaled nitric oxide treatment group and CO represents the placebo gas treatment group. Strata 1 is the 500 to 749 g weight class.................................79

2 Plot of the Moving Average of the P/F Ratio over Time. A 3 window average was used to form the moving averages for Strata 1. CO represents the placebo gas group, and NO represents the inhaled nitric oxide group.................................90

3 Simulation Study Rejection Rates for IUT, Max Test and HT-Test. Equal mean differences refers to how the means were simulated in simulation study, section 1. Colored lines are different combinations of variance and correlation. “Var” represents variance, and “Cor” represents correlation. All results are shown for alpha=0.05..........................................................94

4 Simulation Study Rejection Rates for IUT, Max Test and HT-Test. Real mean differences refers to how the means were simulated in simulation study, section 1. Colored lines are different combinations of variance and correlation. “Var” represents variance, and “Cor” represents correlation. All results are shown for alpha=0.05..........................................................95

5 Simulation Study Rejection Rates for IUT, Max Test and HT-Test. One mean difference refers to how the means were simulated in simulation study, section 1. Colored lines are different combinations of variance and correlation. “Var” represents variance, and “Cor” represents correlation. All results are shown for alpha=0.05..........................................................96
1. **Introduction to Equivalence Tests**

This dissertation follows the three paper format. Within this dissertation we look at equivalence testing from a frequentist point of view and do not delve into the content that includes Bayesian equivalence tests. Equivalence tests, in particular tests of multiple group equivalence and multivariate equivalence have not been fully explored or developed in the existing literature. Tests of equivalence for 2 groups, however, have been studied extensively in bioequivalence studies that are intended to establish the equivalency of two drugs. Thus, before we delve into tests of equivalence for more than 2 groups, or for multi-variables, we first give an explanation of just what is equivalence; and present an overview of the existing literature for equivalence tests.

The definition of equivalence depends on the hypothesis of interest. Depending on the knowledge of the experimenter, the type of study being conducted, and the type of hypothesis being tested what is considered an acceptable limit will vary. The equivalence limit is often set by the FDA based on a set of guidelines that researchers must follow when conducting a clinical trial on the general public (Chow and Liu 2000, Williams et al. 2002). In the late 1960s the FDA began establishing certain regulations for the development and sales of generic drugs. At this time hypothesis tests of equivalence began to garner more attention. Statistical founders such as Fisher, Neyman, and Pearson had already established the null hypothesis test for detecting differences among treatment groups; these tests can be found in most statistical textbooks. Equivalence tests are a form of null hypothesis testing, but although “equivalence” may mean the opposite
of “difference”, the hypothesis test for equivalence is not the direct opposite of a hypothesis test for difference. One cannot merely conclude if we reject the hypothesis of difference among groups that the groups are equivalent. In order to conclude that two or more groups are equivalent we must construct an appropriate hypothesis test that protects against type I error similar to how we would construct a hypothesis test to compare and test for dissimilarity among groups. Early work by Westlake (1981), Anderson and Hauck (1983), and Schuirmann (1987) introduced the average bioequivalence testing hypothesis. This hypothesis can be written as

$$H_0 : |\mu_d| \geq \theta \text{ versus } H_1 : |\mu_d| < \theta$$

where $\mu_d$ represents the population mean difference and $\theta$ is the equivalence limit. The hypothesis stated above in equation (1) is an average equivalence hypothesis since its inference is based on the means. A different form of average equivalence test is the individual equivalence test. The parameters tested in this particular test are related to each individual or subject. Another approach to determine equivalence is the population equivalence test, this approach tests not only for closeness in terms of the mean values, it also tests for closeness in terms of variances. In this paper we shall consider only average equivalence tests; however, the reader may learn more of individual and population equivalence tests (in a bioequivalence framework) by referring to Anderson and Hauck (1990), and Anderson (1993). The test statistic for testing the hypothesis in (1) has many forms. Perhaps the most well-known test associated with the equivalence
hypothesis of (1) is the test that uses “two one-sided tests” known as “TOST”. The steps for performing this test are given in brief below

Step 1: Suppose that $X$ and $Y$ represent a vector of samples from 2 different groups of size $m$ and $n$ respectively. Define $D$ to be the estimator of $\mu_d$ and $S_d$ be the estimator of $\sigma_d$ (the population standard deviation assuming equal variance), and suppose $D \sim N(\mu_d, \sigma_d)$.

Step 2: Calculate $t = (\theta - |D|) / S_d$

Step 3: If $t > T_{\alpha, m+n-2}$ we reject the null hypothesis and conclude the two groups are statistically equivalent. $T_{\alpha, m+n-2}$ is the upper $\alpha$ critical point of a central $T$ distribution with $m+n-2$ degrees of freedom.

This is essentially the test proposed by Schuirmann (1987). It is also an example of a test that uses the intersection union principle (IUP) (Berger 1982). Brown Hwang and Munk (1995) offered an alternative size alpha test to TOST that was uniformly more powerful. The test proposed by Brown, Hwang, and Munk was not considered practical however, due to its irregularly shaped critical region. Berger and Hsu (1996) proposed a “nearly unbiased” test with a more uniform rejection region that they showed to have power similar to the test of Brown, Hwang, and Munk. Brown, Hwang, and Munk (2000) later showed that every unbiased test for the average equivalence hypothesis has an irregularly shaped critical region and gave 3 biased alternatives to the test they proposed in 1995. Of the 3 tests they gave, 2 were modified so that the parameter space was in a sense truncated to make a more intuitive rejection region, however, it appears that in order to make the test more appealing for practice they sacrificed other qualities such as size and unbiasedness of the test. Both the truncated tests were shown to perform poorly when the
variance was large relative to $\mu_d$. The third test, the “S-homogeneous” test was shown to perform relatively well. For a more explicit derivation of these tests please refer to their paper entitled “Testing Average Equivalence-Finding a Compromise Between Theory and Practice” (Brown Hwang and Munk 2000). Recently Wellek (2010) in his book “Testing Statistical Hypothesis of Equivalence and Noninferiority” considered another test for comparing two independent groups; i.e. a non-central t test to test the scaled hypothesis of

$$H_0 : \frac{|\mu_d|}{\sigma_d} \geq \theta \text{ versus } H_1 : \frac{|\mu_d|}{\sigma_d} < \theta$$

(2)

Wellek showed this is the uniformly most powerful (UMP) level alpha test for the average equivalence hypothesis. Wellek’s non-central t test null hypothesis is essentially a test of whether the non-centrality parameter is less than the equivalence limit. The difference in the statement of the hypothesis in (2) with (1) is that the hypothesis statement includes the variance. Intuitively this restatement of the original hypothesis makes sense since in statistics we prefer to represent differences in terms of their number of standard deviations. Besides using a test statistic to test the hypotheses given here, there are other ways of testing these hypotheses such as using confidence intervals.

Additionaly, other equivalence tests exist for 2 group equivalence; these include nonparametric tests, as well as tests for binary data. For a more exhaustive look at equivalence tests please refer to Wellek’s book “Testing Statistical Hypothesis of Equivalence and Noninferiority (2010).
2. Microarray Experiments

Microarray is a high-throughput technology that enables researchers to measure genome wide gene expression, genotype, protein binding, and methylation. A DNA microarray is usually a solid surface to which potentially thousands of DNA molecules are fixed in locations that are known as “probes”. Each of these probes can have several million copies of identical DNA molecules that correspond to a gene. In gene expression assay, a tissue sample is collected from a subject, such as human or mouse, the tissue sample is then put through a process that enables the biologist to extract the mRNA. The mRNA is then hybridized to the microarray to bind with the complimentary DNA. The activity level of the genes represented by the amount of mRNA can be viewed using a substance like a fluorescent dye and a special light source. The amount of fluorescence emitted indicates the expression level of a gene, the brighter the fluorescence, the more mRNA, and the higher the activity of a gene. There are 2 types of DNA microarrays, oligonucleotide arrays and spotted or cDNA arrays (such as BD Biosciences, Standford and others). In the past 10 years technology has advanced and cDNA arrays are now rarely used, the most widely used DNA arrays are oligonucleotide arrays. The oligonucleotide microarray can be further subdivided into short (affymetrix, nimblegen, etc) and long (50-80mer such as produced by agilent, illumina, etc) arrays. Additionally, among the arrays that use fluorescent dyes there are one or two color systems. Oligonucleotide array makers such as Affymetrix utilize the one color system while the 2 color system is used more by cDNA array makers. The number of arrays needed to compare groups in a one color array is two times the number needed in two color arrays, so to save cost some might prefer the 2-color system.
Regardless of the type of array the steps in designing the array and then analyzing the resulting data are similar. The important steps to microarray data analysis start with one or more hypotheses determined by the goals of the study. Once the hypotheses are formulated the design can be addressed. Some of the most important factors in a microarray study design include replication (both biological and technical), randomization, and the statistical design of the microarray experiment itself. As with any design that will be analyzed statistically, proper representation of a population’s variability by means of replication is necessary if we wish to make inferences on that population. Replication is also important to the power of the study. In microarray experiments there are two main types of replication: biological and technical. Biological replication is a measure of the variability within the population but between subjects, whereas technical replication is any replication between bioreplication or between genes.

Biological replication is more relevant to power and the ability to generalize, while technical replication is more relevant to controlling systematic error and the ability to reproduce the microarray experiment. After the number of replicates has been determined and the subjects properly randomized the statistical design of the microarray experiment must be addressed. A few of the core statistical designs for 2-color microarray experiments include the reference design, block design (incomplete or complete), and looping design. These designs all have their advantages and disadvantages (see for example Kerr and Churchill 2001) and the choice of which is better depends on the hypotheses being posed. Experimental design for 1-color microarray experiments is more straightforward because it does not require the level of complexity that 2-color arrays require when determining how best to pair the samples.
and label them on each array. Once the design of the microarray has been determined and experiment carried out, the resulting image is processed before data analysis can begin.

3. Microarray Data Analysis

Data analysis begins with preprocessing the data. Preprocessing includes transforming and normalizing the data. The principle behind transforming microarray data is to essentially put all the data on a more equal footing. The raw intensity scores often span a very large interval of values; by transforming the data we can remove artificial bias and create a dataset of more normally distributed values. One of the most commonly used transformations is to take the logarithm base 2 of the expression values. Another type of transformation is using variance stabilization transformations (Durbin and Rocke 2004).

After transformation there may still exist variability due to nonbiological error. By normalizing the data we can remove this source of error thereby allowing our subsequent tests to detect true difference (or equivalence) in expression unclouded by this extraneous error. There are many methods of normalization, a few of them are loess normalization, quantile normalization (Bolstad et. al. 2003), and methods based on linear regression (Yang et. al. 2001). For a further review of normalization methods see (Quackenbush 2002 and Park et. al. 2003).

Once the data has been appropriately normalized it is ready to be analyzed. As with normalization there exists many ways of analyzing microarray data. The 3 major types of
analysis are class prediction class analysis, and class differentiation. Class prediction seeks to classify samples or similar items such as cancer tumors into known categories or to predict to which category a sample belongs (Somarjai et. al. 2003). Class prediction is an application of discriminant analysis and is often applied in cancer studies (Golub et. al. 1999). Class analysis is an application of cluster analysis. For an overview on the various techniques including hierarchical vs. nonhierarchical clustering methods see Datta and Datta (2003). Sometimes the objective of a microarray experiment is to find out which genes are expressed among two or more treatment groups, in this type of situation we would use class differentiation. Of the three methods given class differentiation is the only method that directly involves statistical hypothesis testing. Equivalence testing is an example of class differentiation. Other examples of class differentiation tests include t-tests, mixed (Wolfinger et. al 2001), linear, and Bayesian models (Baldi and Long 2001).

When considering any high dimensional data analysis one must also consider issues of multiple testing as well as the amount of variability generated due to the sheer amount of data (microarray data can consist of multiple treatment groups represented by thousands and thousands of genes). To account for the large amount of variability in high dimensional datasets shrinkage methods are implemented. Shrinkage enables one to borrow information across the thousands of genes to obtain a better estimate of the variance. For more examples of shrinkage in microarray data analysis see Cui et. al. (2005). To address the multiple testing issues, one can control the number of false positives in the usual manner one might do with low dimensional data by using a correction such as Bonferroni’s method or permutation type tests to control the family
wise error rate (Westfall et al. 2002). A more powerful method for controlling the number of false positive tests is using the false discovery rate (FDR). Some of the commonly used methods that use FDR to control type I error in the microarray setting include Benjamini and Hochberg’s procedure (Benjamini et al. 2001), and John Storey’s q-value (Storey et al. 2003). Other methods include mixture model and Bayesian methods (Do et al. 2005). For a more in depth review of microarray multiple testing issues as well as microarray experimental design and data analysis consider the book by Daniel P. Berrar, Werner Dubitzky and Martin Granzow “A Practical Approach to Microarray Data Analysis”.

4. Equivalence Tests Applied to Microarray Data and gene analysis

The majority of the existing equivalence tests we discussed so far were originally developed for bioequivalence testing. However, these tests can also be applied to genomic data, such as RT-PCR data and microarray data. An instance of equivalence tests being applied to microarray data analysis can be seen in the study published by Qiu and Cui (2010). Qiu and Cui used TOST to identify equivalently expressed genes using real data analysis of two Affymetrix datasets as well as used simulation study based on the real data to evaluate TOST’s performance in terms of power and FDR. Their simulation study showed the power of TOST remained high (almost 100%) with fixed FDR even with a small sample size of 4 when the proportion of nondifferentially expressed genes in their simulation was greater than 0.40 and the equivalence limit no greater than 1.5. Their results seem to indicate that using a fixed FDR cutoff, a typical microarray dataset can achieve high power using TOST with an equivalence limit of no
greater than 1.5. Another microarray study done by Eijgelaar et. al. (2010) used an equivalence method to find equivalently expressed genes across 4 different tissue macrophages. Eijgelaar et. al. used the modified version of TOST presented in Berger and Hsu’s paper (1996). They used confidence intervals rather than a direct test to test for equivalence between each pair of tissues. They combined their equivalence test results with differential gene analysis to aid in pathway discovery and analysis. Finally, equivalence tests have been applied to microarray data where the expression patterns were evaluated over time. Tuke et. al. 2009 used an equivalence confidence interval inclusion approach to identify genes with expression patterns that varied in a particular way across the time condition. An additional application of equivalence tests in genetic data analysis is their use in RT-PCR experiments. In these types of experiments it is necessary to normalize the data to a reference gene. Haller et. al. used equivalence tests to find appropriate reference genes. Their results indicate that there is “no universal reference gene” (Haller et. al. 2009); and they showed that using an equivalence test to detect suitable reference genes can help lead to more accurate analysis in RT-PCR experiments. Similarly, Chang et. al. (2010) applied equivalence tests in their colorectal cancer microarray study. They used RT-PCR to validate the results of their microarray analysis. Like Haller et. al. (2009) they used equivalence tests to find suitable reference genes for their RT-PCR analysis.

5. Multi-group Equivalence Tests

Two group equivalence tests have been applied to microarray data with promising results; however, there are few published studies that have utilized equivalence tests for more than two treatment groups, and to the best of our knowledge multi-group tests have not
been applied to microarray studies in the published literature. Wellek (2010) observed that one of the reasons for the relatively small amount of literature on multi-group equivalence tests is that two group equivalence tests can be extended to k groups (k>2) using the intersection union principle (Berger 1996). In a nutshell, this principle essentially shows that testing all k choose 2 groups for equivalence using a two group level alpha equivalence test such as TOST will result in a level alpha test. The problem with applying the union intersection principle to construct an overall k group equivalence test is that it usually results in an overly conservative test resulting in low power. A simple example of a test that is an application of the UIP is TOST, while TOST is a 2 group test, it utilizes two one sided t-tests. Lauzon and Caffo (2009) actually proposed a multiple group equivalence test based on TOST and a modified bonferroni correction. They showed that scaling α, the type I error rate, by k-1 (k is the number of treatments) provides a method for controlling the family wise error rate that is less conservative than using the Bonferroni correction. The drawback of their proposed method is that it relies on adding the error rates for each comparison. Additionally, it is based on a test (TOST) that has been shown to have low power for detecting equivalence between 2 groups (Brown et. al. 2000), this result will be compounded by the number of pair-wise comparisons it is necessary to test with TOST when considering a k treatment group test. Perhaps a better test for k group equivalence is the F test proposed by Wellek (2010). The F test Wellek proposes is actually the equivalence version of the classical one way ANOVA F test and utilizes the squared Euclidean distance as a summary of the overall distance between k group means. The test he describes is the uniformly most powerful (UMP) invariant test of its kind. Additionally, as it is a modification of the F test, it
shares the same desirable properties such as being an unbiased, level alpha test. We will derive this test in greater detail in our multi-group equivalence paper. The other test we shall consider in greater detail in *Multi-group Equivalence Tests And Their Application to Microarray Data* is the range test for equivalence (Wellek 2010). The range test Wellek proposes is similar to the Studentized range test for differences. It uses the maximum mean difference between group means as the measure of similarity between groups. Since the test uses the maximum mean difference as the measure of similarity, rejecting the null hypothesis in the standard scenario - testing for differences- implies that at least one paired difference between treatment means is significantly different. However, in the equivalence testing scenario rejecting the null hypothesis indicates that all the treatment means must be found significantly equivalent, resulting in a more stringent test in the equivalence testing scenario than the test is meant to be in the usual one way ANOVA scenario. Lastly, Wellek (2010) gives a nonparametric k group equivalence test which is an augmentation of the Mann-Whitney test. While the test has the nice property of requiring no distributional assumptions, it requires a great deal of computation for the formulation of the covariance matrix used in its test statistic and may not be practically useful in microarray data analysis.

6. Multivariate Equivalence Tests

The tests we have discussed in the previous sections are designed for data with only one dependent variable. In this section we shall look at tests of equivalence for data with more than one dependent variable and more than one treatment group, specifically, two group multivariate data with multiple dependent variables. Depending on the context of
the study, the variables in a multivariate dataset may be represented by different quantities. For example, in a microarray study multivariate data might refer to data taken from tissues and several genes where the investigators are interested in testing the equivalence hypothesis simultaneously for all gene expression values between tissue groups. Alternatively, in a bioequivalence study, multivariate data might be data from a control and a treatment group on which “multiple endpoints” such as blood pressure, cholesterol level, etc., have been measured. Regardless of the study context however, all the multivariate equivalence tests discussed in this section are adaptable to any multivariate scenario. In section 6 we showed that the union intersection principle (UIP) can be applied to a single test to create a multi-group test. In a similar manner, the union intersection principle can be used to obtain a multivariate test. Wang et. al. (1999) showed that applying the UIP to test for equivalence using Schuirmann’s two one-sided t-tests (TOST) (Schuirmann 1987) leads to a multivariate test with size alpha less than the nominal for testing equivalence. This test, which we’ll call the intersection union test (IUT) concludes for overall equivalence if for each dependent variable we reject the null hypothesis of nonequivalence between the two groups using TOST. Since it has been shown (Berger and Hsu 1996, Brown et. al 1997) that TOST is not a powerful test and can be improved upon in terms of power, Wang et. al. (1999) proposes a multivariate test for bioequivalence studies which also uses the UIP, but instead of using the one-dimensional TOST they use the one dimensional test of Brown et. al. (1997). This test is uniformly more powerful than the IUT since the one dimensional test of Brown et. al. is uniformly more powerful than TOST (Brown et. al. 1997). A limitation of this modified IUT is that while Brown et. al.’s test is more powerful than TOST, the rejection space of
Brown et. al.’s test has an irregular shape which makes it difficult to specify a meaningful equivalence limit. Additionally, Brown et. al.’s test employs a recursive testing process and is computationally time consuming (Berger and Hsu 1996). These issues suggest that the modified test of Wang et. al. (1999) would not be adequate for application to a high dimensional dataset such as microarray data. Sarandasa and Krishnamoorthy (2005) offer an alternative, the ‘SK’ test, to the IUT and the modified IUT. Their test utilizes results by Halperin (1961) to reduce the multivariate equivalence testing problem to a univariate test. In order for their test to be applicable, however, they make the assumption that all the dependent variables have the same mean difference between comparison groups (Sarandasa and Krishnamoorthy 2005), which for most situations is not a realistic assumption. For example, suppose we measure blood pressure and heart rate and compare these two variables simultaneously between control and treatment groups. SK’s test would require the assumption that the difference between control and treatment group is the same mean constant for both blood pressure and heart rate.

Sarandasa and Krishnamoorthy used simulation study to compare their test and Wang et. al.’s IUT test. Their simulated data consisted of 2 different dissolution profiles and 4 time points. The results of their simulation study indicate that the IUT has lower power than the SK test. However, the power of the IUT remained close to SK’s test for all but a few of the parameter settings, mainly those settings where the “mean shift” (the mean shift is essentially the value represented by the difference between the two dissolution profiles mean vector) values were higher (Sarandasa and Krishnamoorthy (2005)).

Lastly, we look at the multivariate equivalence test given by Wellek (2010). Like the test of Wang et. al. Wellek’s equivalence version of Hotelling’s $T^2$ test for two groups has a
critical region that is somewhat problematic in application. The shape of the critical region is highly dependent on the covariance (Wellek 2010), which is unknown and needs to be estimated. The equivalence form of Hotelling’s $T^2$ test for two groups is still a desirable test nonetheless; as it is the uniformly most powerful invariant (UMPI) size alpha test of its class (Wellek 2010).
MULTI-GROUP EQUIVALENCE TESTS AND THEIR APPLICATION TO MICROARRAY DATA
ABSTRACT

Hypothesis tests of equivalence are typically known for their application in bioequivalence studies and acceptance sampling. Their application to gene expression data, in particular high dimension gene expression data, has only recently been studied, however. In this paper we observe how two multi-group equivalence tests, the F test and the Range test perform when applied to microarray expression data. We adapted these tests to a well known equivalence criteria, the difference ratio (DR). The tests were applied to both single gene and high dimension simulation data. The results show that both tests can achieve moderate power while controlling the type I error at nominal level for typical microarray expression data with the benefit of easy to interpret equivalence limits. For the range of parameters simulated in this paper the power of the F test is higher than that of the Range test, however for a group size of 3 the Range test has similar power to the F test. Finally, the multi-group tests were applied to a real microarray dataset and were used to test for the equivalence of several contrasts simultaneously.
1. INTRODUCTION

1.1 The Equivalence Hypothesis Test

Equivalence analyses have been extensively applied in bioequivalence studies, which are intended to establish the equivalency of two drugs (Chow and Liu 2008, Schall and Luss 1993). The equivalence limit is often set by the FDA based on a set of guidelines that researchers must follow when conducting a clinical trial (Chow and Liu 2000, Williams et al. 2002). In the late 1960s the FDA began establishing certain regulations for the development and sales of generic drugs. At this time hypothesis tests of equivalence began to garner more attention. Statistical founders such as Fisher, Neyman, and Pearson had already established the null hypothesis test for detecting differences among treatment groups; these tests can be found in most statistical textbooks. Equivalence tests are a form of null hypothesis testing, but although “equivalence” may mean the opposite of “difference”, the hypothesis test for equivalence is not the opposite of a hypothesis test for difference. One cannot merely conclude if we reject the hypothesis of difference among groups that the groups are equivalent. In order to conclude that two or more groups are equivalent we must construct an appropriate hypothesis test that protects against type I error similar to how we would construct a hypothesis test to compare and test for dissimilarity among groups. Early work by Westlake (1981), Anderson and Hauck (1983), and Schuirmann (1987) introduced the average bioequivalence testing hypothesis. This hypothesis can be written as

\[ H_0 : |\mu_d| \geq \theta \text{ versus } H_1 : |\mu_d| < \theta \]  

(1)
where $\mu_d$ represents the population mean difference and $\theta$ is the equivalence limit. The hypothesis stated in equation (1) is an *average* equivalence hypothesis since the inference is based on the means. Another form of average equivalence hypothesis is the scaled average equivalence hypothesis given in equation (2). The scaled average equivalence hypothesis is similar to the average equivalence hypothesis, but the population mean difference is scaled by the standard deviation. *equation reference goes here*

$$
H_0: \frac{|\mu_d|}{\sigma_d} \geq \theta \text{ versus } H_1: \frac{|\mu_d|}{\sigma_d} < \theta
$$

(2)

In the book “Testing Statistical Hypotheses of Equivalence and Noninferiority” Wellek considered the noncentral T test for testing the hypothesis in (2), Wellek showed this is the uniformly most powerful (UMP) level alpha test for the scaled average equivalence hypothesis.

A different form of average equivalence test is the *individual* equivalence test. The parameters tested in this particular test are related to each individual or subject. Another approach to determine equivalence is the population equivalence test, which requires both the group means and variances to be within an acceptable limit in order to claim equivalence. In this paper we shall consider only scaled average equivalence tests; however, the reader may learn more of *individual and population* equivalence tests (in a bioequivalence framework) by referring to Anderson and Hauck (1990), and Anderson (1993). The test statistic for testing the hypothesis in equation (1) has many forms.
Perhaps the most well-known test associated with the equivalence hypothesis of (1) is the test that uses “two one-sided tests” known as “TOST”. TOST is an example of a test that uses the intersection union principle (IUP) (Berger 1982).

1.2 Microarray Data Analysis

Microarray is a high-throughput technology that enables researchers to measure genome-wide gene expression, genotype, protein binding, and methylation. A DNA microarray is usually a solid surface to which potentially thousands of DNA molecules are fixed in locations that are known as “probes”. Each of these probes can have several million copies of identical DNA molecules that correspond to a gene. In gene expression assay, a tissue sample is collected from a subject, such as human or mouse, the tissue sample is then put through a process that enables the biologist to extract the mRNA. The mRNA is then hybridized to the microarray to bind with the complimentary DNA. The activity level of the genes represented by the amount of mRNA can be viewed using a substance like a fluorescent dye and a special light source. The amount of fluorescence emitted indicates the expression level of a gene, the brighter the fluorescence, the more mRNA, and the higher the activity of a gene. There are 3 major types of applications for microarray expression data: class prediction class analysis, and class differentiation. Class prediction seeks to classify samples or similar items such as cancer tumors into known categories or to predict to which category a sample belongs (Somarjai et. al. 2003). Class prediction is an application of discriminant analysis and is often applied in cancer studies (Golub et. al. 1999). Class analysis is an application of cluster analysis. For an overview on the various techniques including hierarchical vs. nonhierarchical clustering methods see Datta and Datta (2003). More often the objective of a
microarray experiment is to find out which genes are differentially expressed among two or more treatment groups; in this type of situation we would use class differentiation. Of the three methods given class differentiation is the only method that directly involves statistical hypothesis testing. class differentiation tests include t-tests, linear models (Allison et al. 2006), and Bayesian models (Yang D. et. al. 2004, Gottardo et al 2005).

1.3 The Application of Equivalence Tests to Microarray Data Analysis

Thousands of studies have been published since the arrival of microarray technology; however, only a handful of these articles have focused on the application of equivalence tests to this type of data (Haller et. al. 2009, Tuke et. al. 2009, Chang et. al. 2010, Eijgelaar et. al. 2010, Qiu and Cui 2010). Examples of situations where equivalence tests would be valuable in microarray data analysis include identifying housekeeping genes, identifying common molecular mechanics between diseases, and identifying genes with common expression profiles across treatments or conditions. In a low dimensional setting equivalence tests have been applied to RT-PCR data and been successful in detecting genes that were constantly expressed (see for example Haung et. al. 2006; and Haller et. al. 2004). Additionally, equivalence tests have been applied to microarray data to identify genes with expression patterns that varied in a specific way across the time condition (Tuke et al. 2009). Equivalence tests have also been shown to be useful in identifying common subsets of genes underlining different diseases or between model system and human diseases (Eijgelaar 2010). For comparing 2 treatments, Qiu and Cui (2010) evaluated the performance of TOST in identifying equivalently expressed genes. They used real data analysis and simulations based on the real data to evaluate the performance of TOST in terms of power. Their findings showed that using a reasonable
equivalence limit, TOST can be applied to microarray data and have relative high power while still keeping the FDR low.

Although two group equivalence tests have been applied to microarray data with promising results, there are no published studies involving tests of equivalence applied to microarray data with more than two groups. Wellek (2010) observed that one of the reasons for the relatively small amount of literature on multi-group equivalence tests is that two group equivalence tests can be extended to k groups (k>2) using the intersection union principle (Berger 1982). Simply put, this principle establishes that testing all paired groupings for equivalence using a two group level alpha equivalence test such as TOST will result in a level alpha test. However, the problem with applying the union intersection principle to construct an overall k group equivalence test is that it usually results in an overly conservative test with low power. Lauzon and Caffo (2009) proposed a multiple group equivalence test based on TOST and a modified Bonferroni correction. They showed that scaling $\alpha$, the type I error rate, by $k-1$ ($k$ is the number of treatments) provides a method for controlling the family wise error rate that is less conservative than using the Bonferroni correction. The drawback of their proposed method is that it relies on adding the error rates for each comparison. Additionally, it is based on a test (TOST) that has been shown to have low power for detecting equivalence between 2 groups (Brown et. al. 2000), this result will be compounded by the number of pair-wise comparisons that is necessary to test with TOST when considering a k treatment group test. Perhaps a better test for k group equivalence is the F test proposed by Wellek (2010). The F test Wellek proposes is actually the equivalence version of the classical
one way ANOVA F test and utilizes the squared Euclidean distance as a summary of the overall distance between k group means. The test he describes is the uniformly most powerful (UMP) invariant test of its kind. Additionally, as a modification of the F test, it shares the same desirable properties such as being an unbiased, level alpha test. The other test we will consider in greater detail is the range test for equivalence (Wellek 2010). The range test Wellek proposes is similar to the studentized range test for differences. It uses the maximum mean difference between group means as the measure of similarity between groups. Since the test uses the maximum mean difference as the measure of similarity, rejecting the null hypothesis in the standard scenario - testing for differences- implies that at least one paired difference between treatment means is significantly different. However, in the equivalence testing scenario rejecting the null hypothesis indicates that all the treatment means must be found significantly equivalent, resulting in a more stringent test in the equivalence testing scenario than the test is meant to be in the usual one way ANOVA scenario. In this paper, we evaluate both the range test and F test for their performance on microarray data using simulations and a real data analysis. The tests are applied to data simulated under two different scenarios with a variety of parameter settings. The tests are evaluated in terms of type I error, and power.
2. METHODS

To illustrate the methods we used, we will consider a general situation for a balanced gene expression microarray. Suppose we have genes, \( g=1, \ldots, G \), and for each gene, there are \( k \) treatments. Thus, we have the following data structure: within a given gene, we have observations \( X_{ijg} \), from a \( \text{N}(\mu_{ig}, \sigma^2_g) \), where \( i=1, \ldots, k \) treatments, \( j=1, \ldots, n \) samples, and \( g=1, \ldots, G \) genes. In this setup we assume equal variance across treatments within genes.

2.1 The F Test

We define the overall mean of the \( k \) samples as \( \bar{\mu}_g = \frac{1}{k} \sum_{i=1}^{k} \mu_{ig} \). The F test (Wellek 2010) uses a standardized, squared, global Euclidean distance which we denote \( \psi^2 \). For a single gene this is given by

\[
\psi^2 = \frac{1}{k} \sum_{i=1}^{k} \left( \frac{\mu_{ig} - \bar{\mu}_g}{\sigma^2_g} \right)^2
\]

(3)

Equation 1 is a measure of the overall deviation of the treatment means from \( \bar{\mu}_g \). Thus, for each gene we test the hypothesis:

\[
H : \psi^2_g \geq \varepsilon^2_g \text{ vs. } K : \psi^2_g < \varepsilon^2_g
\]

(4)
with $\varepsilon_g^2 > 0$, the appropriate fixed equivalence limit.

With $N = k*n$, we use $\hat{\psi}_g^2$ given by

$$
\hat{\psi}_g^2 = \frac{\sum_{i=1}^{k} (\bar{X}_{ig} - \bar{X}_{-g})^2}{(N-k)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ijg} - \bar{X}_{ig})^2}, \quad \bar{X}_{ig} = \sum_{j=1}^{n} X_{ijg}, \text{ and } \bar{X}_{-g} = \sum_{i=1}^{k} \bar{X}_{ig}
$$

to estimate the parameter $\psi_g^2$. If we multiply $\hat{\psi}_g^2$ by $n/k$ this statistic has an $F$ distribution with $k-1$, $N-k$ degrees of freedom and noncentrality parameter $n\psi_g^2$.

### 2.2 The Range Test

Unlike the $F$ test, which summarizes the magnitude of the difference among treatment means using an overall Euclidean difference, the range test summarizes the magnitude of the pairwise differences using the maximum difference among all the paired treatment means (Wellek 2010), thus the equivalence limit is easy to specify here. If we wish for all paired differences among $\mu_{1g}, \ldots, \mu_{kg}$ be within some limit $\delta$, we just need to ensure the maximum paired difference is less than $\delta$.

The hypothesis for the range test is given by:

$$
H : \max_{i \neq j} | \mu_{ig} - \mu_{jg} | / \sigma_g \geq \delta \quad \text{vs.} \quad K : \max_{i \neq j} | \mu_{ig} - \mu_{jg} | / \sigma_g \leq \delta
$$

where $\delta$ is the equivalence limit we shall define later. The statistic for testing this hypothesis is obtained from the sample using the following statistic which we denote:

$$
R_{sg} = \frac{\bar{X}_{(k)g} - \bar{X}_{(1)g}}{S_g}
$$
where $\bar{X}_{(i)g}$ is the $i$th ordered (from smallest to largest) treatment mean, $i=1,...,k$. And $S_g$ is the pooled sample variance for gene $g$:

$$S_g = (N - k)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ijg} - \bar{X}_{ig})^2$$

where $N = k \times n$, and $\bar{X}_{ig} = \sum_{j=1}^{n} X_{ijg}$

We apply the range test one gene at a time. Since the distribution of the Studentized range statistic is not easily obtainable, we used Monte Carlo (Sawilowsky et al. 2003) methods to simulate the distribution of $R_{sg}$ and obtain approximate critical points and p-values. The algorithm that was used is given in the simulation section in Appendix A.

### 2.3 The F test - Definition of Equivalence Limit

Both tests that have been described are considered tests for “scaled equivalence”. This is because the hypotheses being tested are formulated for the standardized effect size. Hence we use a moment based criteria called the difference ratio (DR) (FDA 2001) for our definition of the equivalence limit. Under the assumption of equal variance the population difference ratio (DR) is given by (Chow and Liu 2009):

$$DR = \sqrt{\frac{\theta^2 + 2\sigma^2}{2\sigma^2}}$$

where $\theta^2$ represents the squared difference between two groups and $\sigma^2$ is the variance. The formulation of the effect in our hypotheses is similar to equation (6). We use the following result given by Casella and Berger (2001)
\[ \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2 = \frac{1}{2k} \sum_{i=1}^{k} \sum_{j=i+1}^{k} (\mu_{ij} - \mu_{ji})^2 \]  

(9)

and after a little algebra we see that when \( k=2 \)

\[ \psi^2_k = \sum_{i=1}^{k} \frac{(\mu_{ij} - \bar{\mu}_{ij})^2}{\sigma_{i}^2} = \frac{(DR^2 - 1)}{2} \]  

(10)

where \( DR \) is a fixed constant and \( k \), the FDA suggests 1.25, however, we use a range of \( DR \) values in our simulation study to determine what values of \( DR \) are the most practical and meaningful for the microarray scenario. Note that since our application is for multigroup situations, where \( k>2 \), we need a way to relate the formula in (9) to the case where \( k>2 \). Since there are \( k \) choose 2 many pairs of \( \mu_i, \mu_j \) we can extend (10) to the case where \( k>2 \) to give the following equivalence limit for the F test.

\[ \delta^2 = \frac{(k-1)}{2} (DR^2 - 1) \]  

(11)

2.4 The Range test - Definition of Equivalence Limit

As we did with the F test we use the same \( DR \) to establish our equivalence criteria for the Range test. Hence, the range test can be related to the \( DR \) as follows.

\[ R_{sg} = \frac{\bar{X}_{(k)g} - \bar{X}_{(1)g}}{S_{g}} = \sqrt{2(DR^2 - 1)} \]  

(12)

Thus, we define the equivalence limit for the range test to be

\[ \delta = \sqrt{2(DR^2 - 1)} \]  

(13)

2.5 FDR

The false discovery rate (FDR) is a method used to control the error rate that
accompanies an experiment or test that is repeated many times, it controls the error rate of the number of significant tests. Benjamini and Yekutieli (2001) showed that the BH procedure of (Benjamini and Hochberg, 1995) controls the FDR at level less than alpha when “the distribution in each of the composite null hypothesis are stochastically smaller than the null distribution under which each p-value is computed” (Benjamini and Yekutieli (2001)), which is true for the composite hypothesis of the equivalence test. Thus, although the BH method controls at less than nominal alpha for the equivalence hypothesis we will use this procedure to control the error rate generated by the multiple equivalence hypothesis tests we will perform.

3. SIMULATION STUDY

3.1 Simulation Design

A simulation study was carried out to evaluate the tests in terms of type I error and power. Although the F and Range test hypotheses differ in their measure of distance between group means, and cannot be compared directly, they may still be evaluated for their individual performances. To put them on a more comparable footing we used the same simulated data to evaluate both tests.

Data was simulated using 2 different scenario. In the first scenario the data for each treatment group was generated independently from a normal distribution. This data can
be associated with data that one would get from just a single gene. In scheme 2 the data was generated using no distributional assumptions. In order to do this, residuals were created from the real dataset described in section 6; these residuals were used to simulate the data for the high dimension simulations. The manner in which this dataset was simulated is described in more detail in Appendix A.

3.2 Simulation Settings

Single Gene Simulation Parameter Settings:

A total of 10,000 simulations were conducted for each unique parameter setting. For our sample size n, we chose values that would typically be found in a microarray dataset, n=(3, 6, 8, 10, 15). For the group size, k, we used the settings of k=(3, 4, 5). The difference ratio was varied using the settings of DR= (1.25, 1.4, 1.55, 1.7, 1.85). The variance settings were chosen using representative values from a real microarray dataset, \( \sigma^2 (0.04, 0.12, 0.24) \), these respective values represent the 1st, 2nd, and 3rd quartiles of a real microarray dataset (Barger 2008). Means for each treatment group were simulated with the following values a= (0.45, 0.35, 0.25, 0.20, 0.15, 0.10, 0.05, 0). The effect size of the F test is summarized by the square root of the noncentrality parameter (SNCP),

\[
\sqrt{\frac{n \sum_{i=1}^{k} (\mu_i - \mu)^2}{\sigma^2}}
\]

while the effect size of the range test is the standardized maximum mean difference,
“SMax” among any two group means. In order to simulate mean values in a manner where the effect sizes could be calculated, we simulated the means in the following manner:

a) For $k=3$ groups, data is simulated so that $n$ observations come from a $N(a/2, \sigma^2)$ for treatment groups 1 and 2, and $n$ observations from a $N(-a, \sigma^2)$ for treatment group 3.

b) For $k=4$ groups, $n$ observations are simulated from a $N(a, \sigma^2)$ for treatment groups 1 and 2 and $n$ observations are simulated from a $N(-a, \sigma^2)$ for treatment groups 3 and 4.

c) For $k=5$ groups, $n$ observations are simulated from a $N(a, \sigma^2)$ for treatment group 1, $n$ from a $N(a/2, \sigma^2)$ for treatment groups 2 and 3 and $n$ from a $N(-a, \sigma^2)$ for treatment groups 4 and 5.

Thus the square root of the noncentrality parameter (SNCP) for the F test is

$$\sqrt{a^2 / 2\sigma^2}, \sqrt{a^2 / \sigma^2}, \sqrt{7a^2 / 10\sigma^2}$$

when the number of groups, “$k$” is 3, 4, and 5, respectively, for the range test the effect size is $2a/ \sigma$ regardless of $k$.

**High Dimension Simulation Parameter Settings:**

The DR parameter values for scheme 2 include: DR(1.25, 1.4, 1.55, 1.7, 1.85). Mean values, $a(0.45, 0.35, 0.25, 0.20, 0.15, 0.10, 0.05, 0)$ are the same settings as in scheme 1. For scheme 2 the number of replicates are the same as the replicates in the microarray dataset, $n=5$, and $\sigma^2$ is estimated from the subset of data created for scheme 2. The simulations were done for 1000 genes and 500 simulations. For details of how scheme 2 data is generated please refer to “Scheme 2 Simulation” in Appendix A.
3.3. *The Caloric Restriction Mimetic Dataset*

Barger et al (2008) compared the gene expression related to caloric restriction (CR) and resveratrol (RE) supplement in three tissues: heart, muscle, and brain to assess the similarity of molecular mechanisms related to delayed aging resulted from both treatments. Mice were fed with three different diets: a low calorie diet, a normal diet enhanced with resveratrol, and a normal diet to serve as a control. The data are available at GEO with accession number GSE11291.

3.4 *Test for Equivalent Differences for the caloric restriction mimetic data set.*

A key objective of the mimetic study was identifying genes, pathways, and GO terms that are equivalently affected by RE and CR across tissues. Thus, for each gene, we define the means of treatment groups RE and CR as $\mu_{ri}$ and $\mu_{ci}$, respectively, where $i=1,\ldots,3$ represents the different tissues: cortex, muscle, and heart.

To test the equivalence of $\mu_{r1}-\mu_{c1}$, $\mu_{r2}-\mu_{c2}$, and $\mu_{r3}-\mu_{c3}$ we first transform the data as follows:

step 1) Define $y_{rij}$ to be the $j^{th}$ element in the $i^{th}$ tissue of the treatment group RE and $y_{cij}$ to be the $j^{th}$ element in the $i^{th}$ tissue of the treatment group CR.

step 2) Define $m_i=\mu_{ri}-\mu_{ci}$, $i=1, 2, 3$ so that $m_i$ is the mean difference between RE and CR groups for the $i^{th}$ tissue.
Consider the 1-Way ANOVA model constructed from the transformed data:

\[ \mu_i - \mu_{ci} = m_i + \eta_{ij} \text{ where } \eta_{ij} = \varepsilon_{rij} - \varepsilon_{cij}, \]

and \( E(\eta_{jk}) = 0, \text{Var}(\eta_{jk}) = \text{Var}(\varepsilon_{rij}) - \text{Var}(\varepsilon_{cij}) = 2\sigma^2, \ i = 1,\ldots,3, \ j = 1,\ldots,5 \)

with \( \sigma^2 \) estimated by \( S^2 \).

To determine how similar \( \mu_i - \mu_{ci}, \ i = 1,\ldots,3 \) are. We consider testing the following hypotheses using the F and Range test given in the methods section.

hypothesis 1: \( H : \sum_{i=1}^{3} \frac{\bar{m}_i - \bar{m}}{\sigma^2} \geq \varepsilon^2 \quad \text{vs.} \quad K : \sum_{i=1}^{3} \frac{\bar{m}_i - \bar{m}}{\sigma^2} < \varepsilon^2 \)

hypothesis 2: \( H : \max_{h<i} \frac{\bar{m}_h - \bar{m}_i}{\sigma} \geq \delta \quad \text{vs.} \quad K : \max_{h<i} \frac{\bar{m}_h - \bar{m}_i}{\sigma} < \delta; \ h, i = 1,\ldots,3 \)

The noncentral F test is used to test hypothesis 1 and the range test is used to test hypothesis 2. Note in these tests, \( S^2 = 2 \) \( S^2 \) is used.

The procedure we used for finding equivalent differences across tissues can be summarized as follows:

A. Filtration and transformation of genes as described in Barger et al. (2008).

B. Apply ANOVA to determine if there is a significant treatment effect across tissues at nominal 0.05 alpha level.

C. Subtract the mean of CR from the mean of RE for each tissue. Do this for each gene.
D. Gateway step: If the differences are either all positive or all negative continue to the next step.

E. Using this transformed dataset we applied the F and Range test to this data to detect equivalence across the three tissues in both the up regulated (all positive) and down regulated (all negative) gene sets.
4. RESULTS

To evaluate the F and Range tests for their performance on microarray data, we conducted two types of simulations. One is for a single gene under the assumption of normality. The other one has 1000 genes and is based on a real microarray data set. We also applied the tests to a real microarray dataset.

4.1 For Normally Distributed Data the F Test Achieves Good Power While Controlling the type I error.

The non-central F test is a well-known statistic and provides a fairly intuitive method for detecting equivalence among multiple groups, namely, group size greater than two. To evaluate the potential utility of the non-central F test in the equivalence analysis of microarray data, we first conducted some single-gene simulations based on normal distributions for simplicity.

**The power of the F test increases with sample size:**

As expected, when the group sample size increases, the power of the F test increases (Figure 1). For example when sample size is 3, the maximum power that can be achieved is 53% for the variance of 0.12, DR of 1.55 and group size of 4 (panel (e)) in Figure 1). However, the power increases to 86% for sample size of 6 with the same settings of the other parameters. It goes up to 96% and 99% when sample sizes are 8 and 10, respectively.

**The power of the F test increases as the variance increases:**
Because the F test is a scaled hypothesis test, intuitively, increasing the variance (denominator of the effect size in equation 3) reduces the effect size; therefore, it is easier to reject the null hypothesis. This is confirmed by our simulation results (Figure 1). For example, the power increases from 36% to 70% as the variance increases from 0.12 to 0.24 for group size of 4, sample size of 6, and mean value of 0.10 (the orange lines in panels (e) and (f) in Figure 1).

The power of the F test increases along the increase of difference ratio:

For the simulations, we tried different equivalence criterion, which depends on the difference ratio (DR). The results showed that the choice of DR has a major impact on the F test power (Table 1). For a mean value of 0.10, increasing the DR from 1.25 to 1.55 elevates the power from 26.6% to 69.5% for a fixed representative variance of 0.12, group size of 4, and sample size of n=6. Further, increasing DR to 1.85 increases power to 89.6%. The results for the other mean values are similar.

<table>
<thead>
<tr>
<th>Simulated Mean</th>
<th>0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>0.445</td>
<td>0.406</td>
<td>0.266</td>
<td>0.105</td>
<td>0.038</td>
<td>0.008</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.4</td>
<td>0.705</td>
<td>0.667</td>
<td>0.457</td>
<td>0.265</td>
<td>0.126</td>
<td>0.04</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>1.55</td>
<td>0.873</td>
<td>0.818</td>
<td>0.695</td>
<td>0.471</td>
<td>0.246</td>
<td>0.111</td>
<td>0.009</td>
<td>0</td>
</tr>
<tr>
<td>1.7</td>
<td>0.951</td>
<td>0.92</td>
<td>0.834</td>
<td>0.644</td>
<td>0.404</td>
<td>0.225</td>
<td>0.026</td>
<td>0</td>
</tr>
<tr>
<td>1.85</td>
<td>0.972</td>
<td>0.958</td>
<td>0.896</td>
<td>0.747</td>
<td>0.537</td>
<td>0.313</td>
<td>0.051</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 1: Power of the F test as a function of the difference ratio (DR) and simulated mean value. Group size is set as 4; variance is set as 0.12; and sample size is set as 6.
Figure 1: Power curves for the F test from single gene simulations. The columns represent the different group sizes (k=3, 4, 5), and the rows represent different variances ($\sigma^2=0.04$, 0.12, 0.24). The sample sizes (N values) are indicated by the legend in the first panel. The simulated mean value represents the group mean values denoted as “a” throughout the text. The difference ratio (DR) is set as 1.55
The type I error of the F test is on Target:

The type I error of the F test was evaluated using simulations with effect size set on the boundary of the null hypothesis. Table 2 shows the results for DR of 1.55. It is clear that the rejection rate is fluctuating around the target alpha value of 0.05. The fluctuation in the range of values for type I error seems independent of variance, sample or group size, and more likely due to variability within the simulation itself. Similar results were obtained for other DR settings.

Table 2: Type I error for the F and Range tests from Single-Gene Simulations. k, denotes number of groups; and n, denotes sample size. Difference ratio (DR) is set for 1.55. Type I error rate is set as 0.05.
The power of the F test is affected by increasing the number of groups:

To investigate how group number affects the power of the F test we simulated data with mean values that would give the same effect size across groups. To compare power for different group numbers we used the square root of the noncentrality parameter as our measure of effect size: 

$$\sqrt{\frac{n_i \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2}{\sigma^2}}$$

(Liu and Raudenbush 2004). Figure 2 shows the power of the F test for a DR of 1.55, variance of 0.12, and sample size of 6. The power curves for each of the group numbers (3, 4, and 5) show that increasing the number of groups increases the power of this test. For example, for an effect size of 0.05, the power is 0.758 when group size is 3, 0.873 when group size is 4, and 0.924 when group size is 5.
Figure 2: The Power of the F test is affected by number of groups. In this figure the power is given for a difference ratio criterion of 1.55, variance of 0.12, and sample size of 6. Adjusted effect size is calculated as the square root of the noncentrality parameter given in equation (14).

4.1 The Range Test Also Achieves Moderate Power but with conservative Type I error for the Single Gene Simulation Assuming Normality.

Unlike the F test which controls the difference among group means through the noncentrality parameter, the Range test controls the mean difference using the standardized maximum differences among groups. The rejection region of the range test
is more intuitive than that of the F test. To compare the Range and F test on as equal a footing as possible we used the same simulated dataset to evaluate both tests.

In table 2 we observed the type I errors for both tests when the DR is set at 1.55 and alpha at 0.05. The type I error for the Range test is overly conservative for group sizes great than 3. In fact, regardless of variance or sample size there is a definite trend in the decrease in type I error as group size increases.
Figure 3: Power for the Range test from the single-gene simulations. The columns represent the different group size, k=3, 4, 5, and the rows represent different variances $\sigma^2=0.04, 0.12, 0.24$. The sample sizes (N values) are indicated by the legend in the first panel. The simulated mean value represents the group mean values denoted as “a” throughout the text. The difference ratio (DR) is set as 1.55.
The Power of the Range Test Increases with Sample Size:

Similar to the F test, increasing sample size increases the power of the range test (Figure 2). For example, the power expands from 0.153 to 0.364 when sample size is increased from 3 to 6 for a difference ratio of 1.55, variance of 0.12, group size of 4, and simulated mean value of 0.10. Power further expands to 0.577 when sample size is increased to 10.

The Power of the Range Test Increases along with Variance:

Like the F test, the range test for equivalence is a scaled hypothesis test. As such, a larger variance yields an overall smaller effect size, which increases the power to reject the null hypothesis. For illustration, consider the results in Figure 2: the power of the range test increases from 12.4% to 35.7% when the variance is increased from 0.04 to 0.12 for a difference ratio of 1.55, group size of 4, and simulated mean value of 0.10. Power further increases to 43.0% when variance is increased to 0.24 for the same parameter settings.

Increasing DR increases the power of the range test:

The difference ratio greatly affects the Power of the Range test. Table 3 shows that the power increases from 0.022 to 0.156 when the DR is increased from 1.25 to 1.55 and increases to 0.39 when the DR is increased to 1.85 for a representative sample size of 6, variance of 0.12, group size 4, and a simulated mean value of 0.1. This is representative of the results obtained under other parameter setting combinations. These results and
others are given in Table 3.

<table>
<thead>
<tr>
<th>Simulated Mean (columns)/DR (rows)</th>
<th>0</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
<th>0.25</th>
<th>0.35</th>
<th>0.45</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>0.183</td>
<td>0.154</td>
<td>0.085</td>
<td>0.022</td>
<td>0.009</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.4</td>
<td>0.335</td>
<td>0.297</td>
<td>0.18</td>
<td>0.065</td>
<td>0.027</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.55</td>
<td>0.54</td>
<td>0.519</td>
<td>0.341</td>
<td>0.156</td>
<td>0.065</td>
<td>0.014</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.7</td>
<td>0.712</td>
<td>0.692</td>
<td>0.487</td>
<td>0.288</td>
<td>0.143</td>
<td>0.056</td>
<td>0.004</td>
<td>0</td>
</tr>
<tr>
<td>1.85</td>
<td>0.816</td>
<td>0.766</td>
<td>0.607</td>
<td>0.39</td>
<td>0.203</td>
<td>0.105</td>
<td>0.01</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 3**: Power of the range test for as a function of the difference ratio (DR) and the simulated means when the group size, k=4, and variance=0.12.

4.2 *Power of the Range Test Decreases with Increasing k.*

The power of the range test behaves in a similar fashion to that of the F test for most of the parameter settings. However, one of the main differences in the behavior of the range test is that the power of the range test diminishes when the number of groups increases. The decrease in power for each augmentation in group number is more acute when the change is between the lower group numbers. For example, power declines from 0.516 to 0.365 when the number of groups increases from 3 to 4 and experiences a more moderate decline to 0.333 when the number of groups is 5 for a difference ratio of 1.55, variance of 0.12, and sample size of 6; These results and similar ones can be seen in Table 3 and Figure 2.
<table>
<thead>
<tr>
<th>Mean Value</th>
<th>SMax</th>
<th>k=3</th>
<th>k=4</th>
<th>k=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.629</td>
<td>0.561</td>
<td>0.504</td>
</tr>
<tr>
<td>0.05</td>
<td>0.289</td>
<td>0.573</td>
<td>0.468</td>
<td>0.446</td>
</tr>
<tr>
<td>0.1</td>
<td>0.577</td>
<td>0.517</td>
<td>0.364</td>
<td>0.333</td>
</tr>
<tr>
<td>0.15</td>
<td>0.866</td>
<td>0.399</td>
<td>0.163</td>
<td>0.193</td>
</tr>
<tr>
<td>0.2</td>
<td>1.155</td>
<td>0.266</td>
<td>0.07</td>
<td>0.083</td>
</tr>
<tr>
<td>0.25</td>
<td>1.443</td>
<td>0.193</td>
<td>0.021</td>
<td>0.038</td>
</tr>
<tr>
<td>0.35</td>
<td>2.021</td>
<td>0.04</td>
<td>0</td>
<td>0.002</td>
</tr>
<tr>
<td>0.45</td>
<td>2.598</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: Power of the Range test is affected by Group Numbers. The setting of the other parameters are: difference ratio, 1.55; variance, 0.12; and sample size, 6. SMax is the effect size of the range test given in equation 6. The mean value is the value assigned to the group means in the single gene simulations.

4.4 Range Test is a Less Powerful than the F test for Larger Group Sizes.

The power of the range test is within 15% of the F test regardless of the other parameter settings when group size is 3. Figures 1 and 3 illustrate this result. However, the power of the F test is sometimes twice as much as that of the range test for the same parameter settings when the group number is larger than 3. For example the power of the F test is 0.827 which is more than twice the power (0.333) of the range test for a group size of 5, variance of 0.12, sample size of 6, DR of 1.55, and a simulated mean value of 0.10.
5. HIGH DIMENSION SIMULATION RESULTS

In order to study the power of these tests in a more realistic microarray data setting, we used a sample of 1000 genes (more details in the “scheme 2” simulation in Appendix A) from the Caloric Restriction Mimetic dataset, and explored how the power behaved for different values of the means, and difference ratio. The sample and group sizes are set the same as the original data, 5 and 3, respectively. The variance of each gene is estimated from the real data sample. Thus, in this scenario, the power analysis is viewed as more of an “average” power across the genes.

5.1 Power Increases with Increasing DR for Both the Range and F Tests

The difference ratio highly influences power for the multiple gene simulation. As with the single gene simulation, increasing the difference ratio increases power. For example, the power of the F test increases from 0.005 to 0.276 and the power of the range test increases from 0.022 to 0.275 when the difference ratio is increased from 1.25 to 1.85 for a simulated mean of 0.15, variance estimated from the real data, group size of 3, and sample size of 5. Figure 3 illustrates these results and gives the power curves for other values of the difference ratio.

5.2 Power is Very Similar for Both Tests in the High Dimension Simulation Setting

In the high dimension simulation setting the range and F test for equivalence do not significantly differ in terms of power. This occurrence is likely influenced by the fact that the microarray dataset that was used to create the residuals for the simulations has a group size of 3. In the results from single gene simulations given in Figures 1 and 2 we
observed that the power of the range and F test for equivalence are most similar for group size of 3. In the high dimension setting, however, we observe that the performance of these tests in terms of power is nearly the same for all the difference ratio values. Figure 4 shows almost identical power curves for these tests.

**Figure 4**: Power for the Range and F test for the High Dimension Gene Simulations.

Colored lines represent different values of the difference ratio criterion. The “a” values are the simulated mean values.
6. DATA ANALYSIS

6.1 Data Description

It has been shown in mammals that a calorie reduced (CR) diet can slow certain processes involved in aging (Sohal and Weindruch 1996, Lane et al. 2001). These processes include neurophysiologic diseases such as alzheimers (Zhu et al. 1999), parkinson’s (Duan and Mattson 1999), and tumor production (Weindruch and Walford 1988). Many papers have shown that mice on a CR diet live longer (Lee et. al. 1999), and some have reported that mice on a normal diet supplemented with resveratrol also age more slowly (Barger et al. 2008). It isn't known how similar these two treatments are to one another. In order to quantify the extent of similarity between the effects of resveratrol and CR on the aging processes, Barger et al. fed mice aged 30 months three different diets: a low calorie diet, a normal diet enhanced with 4.9mg/kg day of resveratrol, and a normal diet to serve as a control. The effects of these treatments were investigated across heart, skeletal muscle, and brain tissues using gene expression profiling. GO term analysis indicated that resveratrol does have a similar effect on aging as CR. Barger et al. applied a test for difference to compare CR to the control group and RE to the control group in each tissue, they found that both CR and RE have similar effects on gene regulation in the heart, muscle and brain. The test was applied one at a time to each tissue, thus they were not able to find genes that were similarly expressed across all the tissues in a single test. Additionally, since the test being applied were tests for differences, the amount of similarity between CR and RE was not quantifiable. Thus, we’ve applied our multiple equivalence tests, the range, and F test to investigate the amount of similarity between resveratrol and CR. We’ve used the testing for equivalent
contrasts technique described in our methods section.

The dataset we analyzed was an affymatrix mouse array consisting of 3 tissues: cerebral cortex (brain), gastrocnemius (muscle), and heart, and 3 treatment groups: old control-fed, i.e. aged mice that were fed a calorie restricted diet, mice (CR), and old resveratrol-treated mice (RE), and an old-control group (control). Within each tissue there were 5 replicates, for a total of 15 samples in each treatment group.

Using the steps given in Section 3.2 to filter and transform the data, the F test and Range test were applied to the mouse dataset using a variety of DR values, the results of which are given in Table 4. The F test identified 322 down-regulated genes and 233 up-regulated ones; while the range test found 9 down-regulated and 5 up-regulated genes. The equivalence criterion, DR, was set at 1.55. Both tests were completed using a BH corrected alpha of 0.05. The gene lists were further investigated using DAVID (Huang et al. 2008 (a) and (b)). The gene list corresponding to the up regulated genes did not have any significant GO terms. The down-regulated genes, however, did have several significant GO terms. One of the significant GO terms, chromatin modification (GO:0016568), corresponds to chromatin modeling, which is related to the GO terms identified by Barger et al. in their functional analysis of genes that were similarly expressed in both RE and CR. The other significant GO terms represented by a minimum of 10 genes were involved in metabolic processes: nucleobase-containing compound metabolic process (GO:0006139), macromolecule metabolic process (GO:0043170), nitrogen compound metabolic process (GO:0006807), cellular nitrogen compound metabolic process (GO:0034641), cellular macromolecule metabolic process (GO:0044260).
<table>
<thead>
<tr>
<th>DR</th>
<th>1.25</th>
<th>1.4</th>
<th>1.55</th>
<th>1.7</th>
<th>1.85</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Up</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Test</td>
<td>172</td>
<td>312</td>
<td>478</td>
<td>618</td>
<td>701</td>
</tr>
<tr>
<td>Adjusted F Test</td>
<td>1</td>
<td>4</td>
<td>322</td>
<td>580</td>
<td>690</td>
</tr>
<tr>
<td>Range Test</td>
<td>102</td>
<td>179</td>
<td>297</td>
<td>417</td>
<td>536</td>
</tr>
<tr>
<td>Adjusted Range Test</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>287</td>
<td>463</td>
</tr>
<tr>
<td><strong>Down</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Test</td>
<td>123</td>
<td>226</td>
<td>325</td>
<td>385</td>
<td>415</td>
</tr>
<tr>
<td>Adjusted F Test</td>
<td>1</td>
<td>9</td>
<td>233</td>
<td>367</td>
<td>411</td>
</tr>
<tr>
<td>Range Test</td>
<td>65</td>
<td>133</td>
<td>209</td>
<td>298</td>
<td>348</td>
</tr>
<tr>
<td>Adjusted Range Test</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>200</td>
<td>323</td>
</tr>
</tbody>
</table>

Table 4: Equivalently Expressed Genes identified by the F and Range Test. The DR was varied from 1.25 to 1.85. The adjusted alpha was set at 0.05. The counts are given before and after the Benjamini and Hochberg adjustment.
7. DISCUSSION

We have taken a couple of the most promising approaches for multi-group equivalence testing, the F test and the Range test (Wellek 2010) and have evaluated them using simulation study and applied them to a real microarray dataset (Barger et al. 2008) to identify genes with equivalent. The data was simulated using two different scenarios: single gene simulation and high dimension simulation. The tests were evaluated in terms of type I error, and power. Our simulation results indicate that, for the range of parameter settings investigated, the F test performs better than the Range test in detecting equivalently expressed genes in microarray data. These methods were applied to the mouse dataset (Barger et al. 2008) to examine the similarity in the effect of RE and CR. We created a procedure to examine this relationship across heart, skeletal muscle, and brain tissues using the multiple group Range and F test. Our gene expression analysis of the mouse dataset coupled with GO analysis found results that coincide with the finding of Barger et al. (2008)

7.1 The simulated type I error fluctuates around the nominal value.

In our single gene simulations, we observed that the type I error for the F test is on target throughout all the parameter settings while the range test becomes over conservative along with the increase of group number. This behavior is a function of the rejection region of the Range test; the number of pairwise comparisons increases substantially for every increase in group number. While, the type I error of the F test appears to be on target it does fluctuate around alpha. The variability in the type I error is not a function of a particular parameter setting like in the case of the Range test whose type I error
decreases with increasing k.

7.2 The power of the Range test is comparable to the power of the F test when group size is 3.

While the hypotheses of the Range and F test define a very different equivalence criterion, we used the same simulated data and mean values to compare the performance of these tests on as equal a footing as possible. Both the single and multiple gene simulation results indicate that for microarray data, the Range test performs equally as well as the F test for three group comparison. Since the Range test has a more interpretable effect size it might be preferable over the F test in this scenario. However, for comparing more than three groups, the range test tends to be more conservative and have less power than the F test (Figures 1 and 2), which makes it less useful for larger number of groups.

7.3 The Power increases for both the Range and F Test as the Variance Increases.

A concern of using these tests might be that high variability in the data might result in a test that detects equivalence solely due to the high variability in the dataset. If a scaled hypothesis test is applicable, but the investigator wants to eliminate any small effects that are due to high variance, a step could be done where the variables with variance higher than say the 75th percentile, could be removed from consideration in the final analysis. This phenomenon is not so different from the extra small variances in the literature for detecting differentially expressed genes. In this case, variance shrinkage can help to reduce the effect of extra small variances. We expect that variance shrinkage can have the same benefit in the high dimension equivalence testing.
7.4 **Limitations: Results only representative for certain parameter settings.**

A limitation of this paper is that only 1 microarray dataset was used to create the dataset for the multiple gene simulation power study. It might be more informative to have additional microarrays with various group and sample sizes to better understand the performance of these test in a real microarray setting.

7.5 **FDR**

Current FDR estimation procedures such as those developed by Benjamini and Hochberg (1995), and John Storey (2003) were developed for the simple null hypothesis test. These procedures assume the distribution of the p-values under the null hypothesis is uniform, which is not true for the composite null hypothesis of the equivalence test. However, Benjamini and Yekutieli (2001) showed that the Benjamini Hochberg, “BH”, procedure of (Benjamini and Hochberg, 1995) controls the FDR at level less than alpha when “the distribution in each of the composite null hypothesis are stochastically smaller than the null distribution under which each p-value is computed’’ (Benjamini and Yekutieli (2001)), which is true for the composite hypothesis of the equivalence test. Thus, the BH method controls at less than nominal alpha for the equivalence hypothesis. The only downside is that the it is conservative, which decrease the power.
REFERENCES:


Schuirmann, D. J. (1981). On hypothesis testing to determine if the mean of a normal
distribution is contained in a known interval. *Biometrics* 37:617.


APPENDIX A: SIMULATION ALGORITHMS

Single Gene Simulation Procedure:

F test

1. Generate \( n \) observations from a normal distribution according to a), b), or c) of section 3.2. For example, if \( k=3 \) we generate \( n \) observations according to a), if \( k=4 \) we generate \( n \) observations according to b), and if \( k=5 \) we generate \( n \) observations according to c).

2. Compute

\[
\hat{\psi}^2 = \frac{\sum_{i=1}^{k} (\bar{X}_i - \bar{X})^2}{(N-k)^{-1} \sum_{j=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_j)^2}, \text{ where } \bar{X}_i = \sum_{j=1}^{n} X_{ij}, \text{ and } \bar{X} = \sum_{i=1}^{k} \bar{X}_i
\]

3. Do steps 1-2 10,000 times.

4. Calculate how many times \( \hat{\psi}^2 < \left( \frac{k-1}{n} \right) F_{k-1,N-k,\alpha}(n\varepsilon^2) \)

Range Test

1. Using the data generated in step 1. of the F test, order the group means from smallest to largest.

2. Compute \( \hat{R}_s = \frac{\bar{X}_{(k)} - \bar{X}_{(1)}}{S} \), where, \( \bar{X}_{(k)} \) is the largest ordered treatment mean, \( \bar{X}_{(1)} \) is
the smallest ordered group mean, and \( S \) is the ANOVA estimate of variance as given in equation 1 above.

3. Do steps 1. - 2. 1000 times. Simulate the distribution of \( R \), as described in the methods section using the following steps

   i. Generate \( n \) observations for treatment 1 from a \( N(-\delta/2, \sigma^2) \), \( n \) observations each from a \( N(0, \sigma^2) \) for treatments 2,...,\( k-1 \), and \( n \) observations from a \( N(\delta/2, \sigma^2) \) for treatment \( k \).

   ii. Calculate the mean \( \bar{X}^*_i \) and standard deviation \( S^*_i = i=1,...,k \) for each treatment group.

   iii. Order the means.

   iv. Compute \( \hat{R}^*_i = \frac{\bar{X}^*_j - \bar{X}^*_i}{S^*_i} \).

   v. Repeat steps i.- iv. 10,000 times.

   vi. Order the \( R_i \)' from smallest to largest.

4. The sample 5\(^{th} \) percentile is \( \hat{R}^*_i(500) \) which is an estimate of \( R_{s(0.05)} \).

High Dimension Simulation Procedure:

F test

1. Let \( i=1,...,3 \) be the number of treatments, \( j=1,...,5 \) be the number of replications in a given treatment, and \( g=1,...,1000 \) the number of genes. Using a subset of the public
dataset GSE11291 described in section 3.3 (Barger et. al. 2008) we created a baseline dataset by doing the following: for each observation, $Y_{ijg}$, from the real dataset, subtract off the $i^{th}$ treatment mean, $Y_{ijg} - \overline{Y}_{i,g}$.

2. Randomly select 1000 genes from this baseline dataset.

3. Keeping the order of the genes, but permuting the observations and treatment groups, randomly select $n=6$ observations to be in each of the $k=3$ treatment groups.

4. Add the effect sizes “a” according to section 3.2’s part b) for $k=3$ groups.

5. Calculate $\hat{\psi}_g^2 = \frac{\sum_{i=1}^{k} (\overline{X}_{ig} - \overline{X}_{..g})^2}{(N-k)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ijg} - \overline{X}_{ig})^2}$

6. For $i=1,\ldots,1000$ genes count how many times $\hat{\psi}_g^2 < F_{k-1,N-k,\alpha}(n\epsilon_g^2)$ and divide by 1000. Define this count as $P_r$.

7. Do step’s 3. – 6. 500 times.

8. Calculate the “average power” by averaging the $P_r$’s, $r=1,\ldots,500$. 
Range test

1. Let $i=1,\ldots,3$ be the number of treatments, and $j=1,\ldots,5$ be the number of replications in a given treatment, and $g=1,\ldots,1000$ be the number of genes. Using a subset of the public dataset described in section 3.3 (Barger et. al. 2008) we created a baseline dataset by doing the following: for each observation, $Y_{ijg}$, from the real dataset, subtract off the $i^{\text{th}}$ treatment mean, $Y_{ijg} - \bar{Y}_{i,g}$.

2. Randomly select 1000 genes from this baseline dataset.

3. Keeping the order of the genes, within each gene we permute the observations and treatment groups, and then randomly select $n=6$ observations to be in each of the $k=3$ treatment groups.

4. Calculate $\hat{R}_{sg} = \frac{\bar{X}_{(k)g} - \bar{X}_{(l)g}}{\hat{S}_g}$.

5. Simulate the distribution of $R_{sg}$ using the following steps:
   a) Generate $n$ observations for treatment 1 from a $N(-\delta/2, \sigma_g^2)$, $n$ observations from $N(0, \sigma_g^2)$ for treatment 2, and $n$ observations from $N(\delta/2, \sigma_g^2)$ for treatment 3.
   b) Calculate the mean $\bar{X}_{(i)g}^*$ and standard deviation $S_{(i)g}^*$, $i=1,\ldots,k=3$ for each treatment group.
   c) Order the means.
   d) Compute $\hat{R}_{sg}^* = \frac{\bar{X}_{(k)g}^* - \bar{X}_{(l)g}^*}{\hat{S}_g}$.
   e) Repeat steps a)-d) 10,000 times.
f) Order the $\hat{R}_{sg}^*$'s, from smallest to largest $\hat{R}_{sg}^*(1) < \hat{R}_{sg}^*(2) < ... < \hat{R}_{sg}^*(10,000)$

g) The sample 5th percentile is $\hat{R}_{sg}^*(500)$ which is an estimate of $R_{s(0.05)}$.

6. Count how many times $\hat{R}_{sg} < \hat{R}_{sg}^*(500)$ and divide this count by 1000. We denote this value $Q_i$.

7. Do steps 3. – 7. 500 times.

8. Compute the average power by averaging $Q_i$, $i=1,...,500$. 
EQUIVMULTI: A MULTI-GROUP EQUIVALENCE TEST PACKAGE
Abstract

Microarray is routinely used in many biological experiments to answer questions regarding genome-wide gene expression. In many situations, such as identifying housekeeping genes, identifying genes that have common molecular mechanics or expression profiles the standard test for differences are not applicable. EquivMulti is a package that enables investigators to apply tests of equivalence to both low and high dimensional data where the number of groups is greater than or equal to 2. Each equivalence test within the package was tested on a single gene first, and then used on high dimensional data containing tens of thousands of genes. The functions in EquivMulti have been tested and verified on the low-dimensional dataset given in *Testing for Statistical Hypothesis Tests of Equivalence and Noninferiority*, and applied to the a mouse microarray dataset (GSE1148) available on GEO. Both of these datasets are included in the package as an example.

Availability and Implementation

EquivMulti will be available for download via CRAN.

Contact: ctfyang@gmail.com
Introduction:

Microarray is a high-throughput technology that enables investigators to measure genome-wide gene expression, genotype associations, protein binding, and methylation. There are standard testing procedures to identify differentially expressed genes across treatment or tissue groups using hypothesis tests such as the T test or ANOVA. In many situations, however, it is of interest to the investigator to identify expression patterns of genes that instead of being different are the same across treatment groups or conditions. In such a situation a standard hypothesis test for differences is not applicable as it will not control the type I error. In order to declare statistical equality, and still control the type I error, a hypothesis test of equivalence should be applied. There are many packages that will perform the standard statistical hypotheses tests for differences on microarray data (Churchill et al. 2009). *Equivalence* is the only R package available on CRAN or BIOCONDUCTOR which allows a researcher to perform tests of equivalence on two groups from a normally distributed population (Robinson 2010). Our R package, EquivMulti, allows researchers to execute the multi-group F-test and Range test for testing equivalence across multiple groups as given in (Wellek 2010) with the adaptations outlined in (Yang thesis 2013). *EquivMulti* takes as input a simple dataframe where each column consists of a treatment group, and outputs results in a simple list form. These results contain the test-statistic and effect size for the relevant multi-group test and corresponding p-values.
Method:

EquivMulti uses a simple dataframe as the main input where the columns consist of the number of groups and the rows consist of the number of replicates. A dataset should be inputted using the format:

\[
\begin{pmatrix}
X_{111} & X_{112} & \cdots & X_{11n_k} & X_{121} & X_{122} & \cdots & X_{12n_k} & \cdots & X_{1k1} & X_{1k2} & \cdots & X_{1kn_k} \\
X_{211} & X_{212} & \cdots & X_{21n_k} & X_{221} & X_{222} & \cdots & X_{22n_k} & \cdots & X_{2k1} & X_{2k2} & \cdots & X_{2kn_k} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\
X_{g11} & X_{g12} & \cdots & X_{1n_k} & X_{g21} & X_{g22} & \cdots & X_{g2n_k} & \cdots & X_{gk1} & X_{gk2} & \cdots & X_{gkn_k}
\end{pmatrix}
\]

where \(X_{gkn_k}\) represents the \(n_k\)th subject for the \(g\)th gene in the \(k\)th treatment group.

The F-test of equivalence in the package can handle unequal sample sizes in the dataframe, while the Range test function only allows samples with equal replications.

For the F-test researchers supply arguments such as the significance level (\(\alpha\)), and the difference ratio (DR) criterion (Yang thesis 2013). The notion of DR is a user specified argument and is typically determined beforehand by the investigator. For the Range test researchers will first generate the distribution of the Range statistic via monte carlo simulation where they specify sample size, the DR criterion.

Once arguments are specified, EquivMulti performs the specified Range test or F-test. Performance of the F-test is optimized by using the R built in function \textit{apply} to speed computation through the dataframe (R Development Core Team 2011). Performance of the Range test is similarly optimized. Monte Carlo simulation for the Range test is carried out separately from the main function for computational speed. Using R’s \textit{proc.time} with 22690 genes to analyze, performing both R and F-test, and the \textit{monte}..
Algorithm with R=10000 took 48.25 seconds. This was recorded on an AMD II X4 965 Processor 3.40 GHz with 8 GB ram.

\textit{EquivMulti} was applied to the Affymetrix Mouse Expression 430A array (Rinn et al., 2004) to aid in preliminary dissertation research. To help researchers test the \textit{EquivMulti} package, the data from the array was put into the package for use. This array data is unaltered, and a preliminary application of normalize.qspline() from the affy package, followed by the standard log$_2$ transformation to the result is suggested (Gautier, L., et al. 2004). The design file for this experiment is also included within the package as an example. No functions within \textit{EquivMulti} are dependent upon it. Another dataset used to help validate calculations within the \textit{EquivMulti} package is the diastolic blood pressure dataset from Chapter 7 of \textit{Testing Statistical Hypotheses of Equivalence and Noninferiority} (Wellek 2010). This data comes from a four arm clinical trial where averages of diastolic blood pressure were recorded several times per treatment arm. The intention of including this dataset is that it allows one to test the basic F-test function within the package against Wellek’s work to ensure the accuracy of the function itself.

\textbf{Application}

\textit{EquivMulti} can easily perform multi-group equivalence testing for the researcher. By utilizing the functional approach, a simple call is made to perform the F-test or Range test of equivalence. The return object is a simple list. Making the access of the equivalence tests simple functions allows the user to embed them in further objects, or wrap them in a bigger function for further analyses.
Results

EquivMulti was tested on two main sources of data. The first source of data came from a comparative clinical trial of 4 different antihypertensive treatments where blood pressure was recorded several times for each treatment (Wellek 2010). Computing the F-test statistic on the diastolic dataset gives $\psi^2=0.02222164$, which is precisely what Wellek provides to the 4th decimal position.

EquivMulti was applied to the mouse data using it’s built in F and Range tests of equivalence. In this situation we compared, hypothalamus, liver, kidney, and testis for the male population of mice for each gene. Results for the testing are given below were the number of significant findings at the $\alpha=.05$ level are found before and after FDR adjustment.

<table>
<thead>
<tr>
<th>DR</th>
<th>1.25</th>
<th>1.4</th>
<th>1.55</th>
<th>1.7</th>
<th>1.85</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F-test</strong>: before adjustment / after adjustment</td>
<td>422/0</td>
<td>992/0</td>
<td>1750/38</td>
<td>2594/709</td>
<td>3424/1713</td>
<td>4208/2703</td>
</tr>
<tr>
<td><strong>Range test</strong>: before adjustment / after adjustment</td>
<td>240/0</td>
<td>541/0</td>
<td>987/2</td>
<td>1541/5</td>
<td>2161/251</td>
<td>2869/935</td>
</tr>
</tbody>
</table>

*Table 1*: The number of equivalent genes before and after FDR adjustment using the Benjamini and Hochberg method. DR is the difference ratio criterion given in (Yang 2013).
References


Andrew Robinson and the R Development Core Team (2010). equivalence: Provides tests and graphics for assessing tests of equivalence. R package version 0.5.6.


MULTIVARIATE EQUIVALENCE TESTS WITH APPLICATION TO CLINICAL TRIAL DATA
ABSTRACT

Inhaled nitric oxide therapy for term and near-term infants is commonly used to treat respiratory illness. In this paper we derive a multivariate equivalence test, Max Test, that can be used to evaluate the equivalence of two treatment groups across multiple outcomes or conditions simultaneously. Max Test, and two other standard multivariate equivalence tests the Intersection Union test and Hotelling’s $T^2$ test (the IUT and) have been evaluated in terms of type I error and power using simulated data modeled after the Inhaled nitric oxide data set. The tests have also been applied to the real data and a moving average of the data to test for equivalence of the P/F ratio over 8 time measurements.
INTRODUCTION

1. Inhaled Nitric Oxide (iNO) Therapy in Neonates

Neonatal lung disease is the most common complication of preterm delivery and causes significant morbidity and mortality. Disorders of prematurity and respiratory distress are among the leading causes of infant mortality in the U.S. The use of inhaled nitric oxide (iNO) has been approved by the U.S. Food and Drug Administration for respiratory failure of the term and near-term infant and is recommended by professional societies, such as the American Academy of Pediatrics. Evidence supporting the use of iNO in term or near term infants is summarized by a Cochrane review (2010) that found that use of iNO therapy reduced the need for extracorporeal membrane oxygenation in term and near term (> 34 weeks gestation) infants with respiratory failure. The aim of iNO is to improve oxygen delivery and lung health. There are, however, complications. Nitric oxide when combined with oxygen produces nitrogen dioxide (NO₂), which is a toxic gas. It can also counteract other treatments for related problems, e.g. right to left shunt operations which are needed in neonates who have congenital cardiac anomalies. Thus, if iNO is equivalent to using carbon monoxide or another standard treatment in treating respiratory failure, it is not necessary to submit this already high risk group to treatment that has potentially serious side effects.

Respiratory failure is commonly measured by the oxygenation index (Field et al.1996), recent studies (McVey et al. 2011, McCrae et al. 2004) considered the use of the P/F ratio in an iNO study of preterm infants. The P/F ratio is the ratio of arterial oxygen
The concentration to the fraction of inspired oxygen (PaO$_2$/FiO$_2$). It reflects how well the lungs absorb oxygen from expired air. A higher ratio indicates better gas exchange. One of the purposes of this study is to use a statistical test for equivalence to assess the difference between groups in the P/F ratio over time. If the difference of the P/F ratio between the iNO and CO treatment groups is within a specified limit that is considered clinically equivalent, say no more than 15-20% of the baseline value, then one might conclude there is not a clinically significant difference in oxygenation improvement between iNO and CO groups.

2. Multivariate Equivalence Tests

In order to examine the equivalence of iNO and CO treatments over time, we reviewed the literature for an appropriate multivariable equivalence test. At present there does not appear to be a satisfactory multivariate test for equivalence between two treatment groups. Wang et. al. (1999) showed that applying the Union Intersection Principle, (UIP) to test for equivalence using Schuirmann’s two one-sided t-tests (TOST) (Schuirmann 1987) leads to a multivariate test with size alpha less than the nominal for testing equivalence. This test, which we will call the intersection union test (IUT) concludes for overall equivalence if for each dependent variable we reject the null hypothesis of nonequivalence between the two groups using TOST. In this paper we will apply two different multivariate equivalence tests to the iNO dataset, the intersection union test, “IUT”, of Wang et. al. (1999), and Hotelling’s T$^2$ equivalence test, “HTE” for 2 groups (Wellek 2010) to the iNO dataset. Each of the aforementioned tests has a drawback;
in the case of the latter, the shape of the critical region is difficult to work with in terms of the natural parameter space used for determining distance, in the case of the former it has been shown to have relatively low power (Berger and Hsu 1996, Sarandasa and Krishnamoorthy 2005), which decreases as the variance increases (Ennis and Ennis 2009). Thus, we propose an alternative test, Max test, to the IUT and HTE tests of equivalence, and evaluate the performance of these tests in the hope that Max Test will improve the power and functionality of the two available tests. These tests have been applied to the both the iNO dataset and simulated data.
METHODS

1. Data Structure and Assumptions

For the tests discussed in this section we will refer to data with the following structure:

Suppose we have the following two independent samples:

\[
\begin{align*}
\left( X_{is}, \ldots, X_{i\rho s} \right) &\sim N(\underline{\mu}, \Sigma), \\
\left( Y_{it}, \ldots, Y_{it\rho} \right) &\sim N(\underline{\nu}, \Sigma),
\end{align*}
\]

with 
\[
\begin{align*}
\underline{\mu} = (\mu_1, \ldots, \mu_\rho), \\
\underline{\nu} = (\nu_1, \ldots, \nu_\rho)
\end{align*}
\]

represent the vector of means for variables \(i=1,\ldots,\rho\) in each group and 
\[
\Sigma = 
\begin{pmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1\kappa} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2\kappa} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk}
\end{pmatrix}
\]

so that \(\sigma_{ij}\) is the \(ij^{th}\) element of \(\Sigma\). We define the sample mean vectors and pooled covariance matrix to be

\[
\begin{align*}
\overline{X} &= m^{-1} \sum_{s=1}^{m} \left( X_{1s}, \ldots, X_{\rho s} \right), \\
\overline{Y} &= n^{-1} \sum_{t=1}^{n} \left( Y_{1t}, \ldots, Y_{\rho t} \right)
\end{align*}
\]

and

\[
S = \frac{1}{m+n-2} \left( \sum_{s=1}^{m} \left( X_{1s} - \overline{X}, \ldots, X_{\rho s} - \overline{X}_\rho \right) \left( X_{1s} - \overline{X}, \ldots, X_{\rho s} - \overline{X}_\rho \right)^{\prime} + \sum_{t=1}^{n} \left( Y_{1t} - \overline{Y}, \ldots, Y_{\rho t} - \overline{Y}_\rho \right) \left( Y_{1t} - \overline{Y}, \ldots, Y_{\rho t} - \overline{Y}_\rho \right)^{\prime} \right)
\]

respectively. Hence

\(S\) is the sample estimate of \(\Sigma\) with \(s_{ij}\) the \(ij^{th}\) element of \(S\).
2. Hotelling’s $T^2$ Equivalence Test

_Notion of “Closeness”_

Essential to equivalence testing and how we define our equivalence limit on the parameter space is the notion of “closeness”. In determining whether a set of parameters is equivalent we must summarize how close they are using some measure of distance. There are many ways to measure of distance. The most common measure of distance is the Euclidean distance. The Euclidean distance is simply the measure of distance from point $a$ to point $b$ or in the case of the parameters described above it is simply the distance between the mean vectors $\mu = (\mu_1, \ldots, \mu_p)$, $\nu = (\nu_1, \ldots, \nu_p)$. HTE test for two groups uses the Mahalanobis distance to summarize closeness. The Mahalanobis distance incorporates the covariance of the parameters into its measure of distance. In terms of our parameters space it is defined as

$$
(\mu - \nu)^T \Sigma^{-1} (\mu - \nu)
$$

Each of the hypothesis tests given in the subsequent sections is defined by the measure of distance stated in the hypothesis itself.

_Hotelling’s $T^2$ Test_

Keeping in mind the data structure given above, the hypothesis using the Malahnobis measure of distance is given by

$$
H : (\mu - \nu)^T \Sigma^{-1} (\mu - \nu) \geq \delta^2 \ vs \ K : (\mu - \nu)^T \Sigma^{-1} (\mu - \nu) < \delta^2
$$
where \( \varepsilon^2 \) is the equivalence limit.

The test statistic for testing the hypothesis in 6.4.1 above has the form

\[
T^2 = \frac{mn}{m+n} (\bar{X} - \bar{Y}) S^{-1} (\bar{X} - \bar{Y})'
\]

\( T^2 \) has a noncentral F distribution with noncentrality parameter \( \lambda = (\mu - \nu) \Sigma^{-1} (\mu - \nu)' \)

and \( \rho \) degrees of freedom in the numerator and \( n+m-\rho-1 \) degrees of freedom in the denominator (Anderson 1984, Wellek 2010). The rejection region given below is an exact \( \alpha \) level unbiased test (Wellek 2010).

\[
\{ T^2 < \left( k(m+n-2)/ (m+n-\rho-1) \right) F_{\rho,m+n-\rho-1,\alpha} (m\varepsilon^2 / (m+n)) \}
\]

3. The Intersection Union Test

We have described the data structure and the multivariate Hotellings T2 test for equivalence, in this section we will look at the test proposed by Wang et. al. (1999), the IUT. This test is essentially a univariate test applied \( \rho \) times. In the case of the IUT test the univariate test being applied is TOST where the measure of distance is simply the Euclidean distance between the mean vectors.

Define \( W_i = |\bar{X}_i - \bar{Y}_i| < \delta - t_{n+m-2}(\alpha) \left( \frac{s_i^2}{n+m-2} \right)^{\frac{1}{2}} \) \( i=1,\ldots,\rho \), where \( \bar{X}_i - \bar{Y}_i \) is the \( i \)th sample mean difference, \( \delta \) is the equivalence limit, \( t_{n+m-2}(\alpha) \) is the (1-\( \alpha \))% percentile of the
Student’s t distribution with n+m-2 degrees of freedom, and $s_{ij}$ is the $ij$th element of $S$. Lastly define $\theta_i = \mu_i - \nu_i$. We wish to test the ensuing hypothesis

$$H: \max_i |\theta_i| \geq \delta \text{ vs. } K: \max_i |\theta_i| < \delta$$

The IUT of Wang et. al. (1999) rejects the null hypothesis $H_0$ in favor of $H_1$ if we can conclude the following:

$$W^I = \bigcap_{i=1}^{\rho} W_i = \left\{ \left| \bar{X}_i - \bar{Y}_i \right| < \delta - t_{n+m-2}(\alpha) \left( \frac{S_i^2}{n+m-2} \right)^{1/2}, \forall i = 1, \ldots, \rho \right\}$$

Since $W_i$ is a level $\alpha$ test, the intersection union test, $W^I$, will have level less than $\alpha$ (Wang et. al. 1999 and for a proof of this using the intersection union principle (UIP) see Berger and Hsu 1996), but how much less than $\alpha$? If we assume the variables are highly correlated, then the IUT will have level $\alpha$ close to the level of TOST (or whatever univariate test that we choose to use in combination with the UIP), but if the variables are not highly correlated, the test will have size much less than $\alpha$. Moreover, if the variables are independent of one another, each test $W_i$ would be independent of one another and supposing each test $W_i$ has level $\alpha \approx 0.05$, this would imply that for $\rho$ independent tests, the $\alpha$ level for testing $W^I$ is $\approx 0.05^\rho$. Thus the power of the IUT test or any test which utilizes the intersection union principle is also affected by the amount of correlation among the variables being considered. Ennis and Ennis (2009) demonstrated this effect in their discussion of TOST, which is a test based on the application of the UIP. This result is also illustrated by the simulation results of Sarandasa and Krishnamoorthy (2005); their simulation results show the power of the IUT increases as the correlation
setting increases.

4. Max Test

Our final multivariate equivalence method that we propose to evaluate is a test that uses the maximum mean, i.e. the maximum distance between the mean vectors $\mu$ and $\nu$ distance for its measure of closeness. Consider the hypothesis:

$$H : \max_i |\theta_i| \geq \delta \text{ vs. } K : \max_i |\theta_i| < \delta$$

We define the following rejection rule:

Reject the null hypothesis $H$ if

$$M = \max_{1 \leq i \leq p} \frac{|X_i - \bar{Y}|}{s_{ii} + \frac{s_{ii}}{m + n}}$$

is small, i.e. all the $\frac{|X_i - \bar{Y}|}{s_{ii} + \frac{s_{ii}}{m + n}}$'s are small. To determine how small is small enough, we need to determine the distribution, or more specifically, the percentiles of $M$ under $H$.

We will use a parametric bootstrap to estimate the percentiles of our test statistic. In this manner we will be deriving tests controlling alpha at exactly or less than nominal level, which should improve the power of this test over all the aforementioned tests.

Since $H$ is composite, we need to specify parameter values in $H$ that are as close as possible to $K$. We use the steps given in Simulation Study : 2. Settings to estimate the alpha percentile of $M$. 
5. Summary

The methods presented in the previous sections will be evaluated using a simulation study based on a real multivariate clinical trial dataset. The tests are summarized in the table below.

<table>
<thead>
<tr>
<th>Test</th>
<th>Assumptions</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>IUT</td>
<td>Data is normally distributed.</td>
<td>Straightforward, widely used test.</td>
<td>Does not factor in correlation.</td>
</tr>
<tr>
<td></td>
<td>Equal variance among groups assumption.</td>
<td>Easy to define the equivalence region using the usual Euclidean measure of distance.</td>
<td>Based on a univariate test that has low power.</td>
</tr>
<tr>
<td>HTE test</td>
<td>Data is normally distributed.</td>
<td>UMP invariant test for its class of transformations.</td>
<td>Distance is measured in terms of Mahalanobis distance, which is dependent on the variance-covariance matrix – difficult to define an equivalence limit for the critical region.</td>
</tr>
<tr>
<td></td>
<td>Equal variance.</td>
<td>Robust against violations to normality and equal variance assumption.</td>
<td></td>
</tr>
<tr>
<td>Max test</td>
<td>Equal variance.</td>
<td>Measure of distance in this test is easy to apply and understand. If the time point with the largest distance between groups is found to be equivalent all the other pairs of time points will be considered equivalent.</td>
<td>Because it relies on the maximum distance for its decision for equivalence, this test may be less powerful depending on the number of variables to be compared.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uses simulation to find the distribution of the test statistic.</td>
</tr>
</tbody>
</table>

Table 1: Summary of Multivariate Equivalence Tests Investigated in this Paper.
PRACTICAL APPLICATION

The dataset we will use to specify our simulation settings as well as use for real data analysis is a multi-center clinical trial dataset (Cutter 2006) involving neonates. The infants were randomized to be given either inhaled nitric oxide (iNO) or placebo gas for 21 days or until extubation was performed. Patients were stratified according to 3 different weight classes represented by strata 1: 500 to 749 g, strata 2: 750 to 999 g, and strata 3: 1000 to 1250g. This paper uses a subset of the iNO clinical data evaluated over time for both analysis and simulation settings. The dataset consists of the P/F ratio for iNO (NO) and non-iNO treatment groups (CO) measured over 8 time points within strata 1, the lowest weight group. These measurements do not include the baseline value which we reserved for calculating the equivalence limit and which was approximately 202 for both treatment and placebo groups. After missing data was removed from strata 1 there were n=73 subjects in the iNO group and m=54 subjects in the non-iNO group.

![Mean P/F Ratio Over Time Strata 1](image)

**Figure 1:** Mean P/F Ratio Values Over Time by Treatment Group. NO represents the inhaled nitric oxide treatment group and CO represents the placebo gas treatment group. Strata 1 is the 500 to 749 g weight class.
SIMULATION STUDY

Using parameter settings observed from the real dataset, a simulation study was carried out to evaluate the Intersection Union test, Max test, and Hotelling’s $T^2$ test for power and type I error. The data was simulated for 10,000 replications per unique simulation setting and alpha was set at 0.05.

1. Settings

_Determination of Equivalence Limit_

Frequently, in studies involving iNO for pre-term and low birth-weight newborns, the P/F ratio is used as a marker for improved oxygenation. A P/F ratio of <200 is classified as acute respiratory distress syndrome (ARDS). Often, if the P/F ratio has increased 15%-20% from the baseline pre iNO value, the consensus is that that the iNO or other oxygenation therapy has improved the process of providing oxygen (Baldauf et al. 2001, Macrae et al. 2004, Barton 1998, Dellinger 1998). Conversely, a P/F ratio with improvement of <15-20% from baseline is often considered in the clinical setting to indicate a nonresponse to iNO. Thus, for our equivalence limit we will use this same measure to determine equivalence between iNO and non-iNO treatment groups. The reasoning is that if the mean difference between treatment groups over time is less than 15-20% of the non-iNO value then the difference between the two treatment groups does not represent a clinically significant amount.
*Intersection Union Test and Bootstrap Method Parameter Settings*

In order to investigate how the intersection union test (IUT) and Max Test (MT) compared in terms of power and type I error we used values for the mean difference observed from the real dataset. Multivariate data was simulated following the data structure given in section 1 of Methods. The data was simulated so that there were 8 time points, thus $p=8$. Values of the mean difference between NO and CO groups may follow many different patterns across the 8 time points. To better understand the behavior of how the tests performed on different types of mean difference patterns we chose representative patterns to get a flavor for how these test perform in 3 different situations:

1. **Equal Mean Differences**: The 8 time points have the same mean difference $M$ across all time points.

2. **Real Mean Differences**: The 8 time points have mean difference values estimated from the real dataset ($11.96, 5.61, 2.73, 0.07, 6.97, 13.47, 24.13, M$), and only the first time point $M$ is varied, if $M$ is less than the mean difference at any other time point, we replace that value by $M$.

3. **One Mean Difference**: all the time points have 0 mean with the exception of the first time point $M$.

$M$ is varied according to the values ($50, 48.5, 36, 30, 25, 20, 15, 10, 5, 0$) that are
representative of differences found in the iNO dataset.

The variance-covariance matrix was simulated using the real data variance-covariance matrix settings as well as low, medium, and high variance settings with the assumption of equal variances, variance\(^2\)(1000, 6000, 12000). Likewise, the correlation settings were varied under the assumption of equal correlation using low, medium, and high values representative of the real data correlation=(0.05, 0.4, 0.7). Sample sizes were simulated so that the non-iNO group has a sample size of m=54 and the iNO group has a sample size of m=73. For the equivalence limit we used 20\% of the baseline P/F ratio value measured from the non-iNO group, thus \(\delta=40\).

Setup for Hotellings T2 Test

We used the same simulated data from the IUT and BM setup. However, for the equivalence limit \(\varepsilon^2\) we used 0.5. This value was chosen with the reasoning that if the data were not correlated and the number of variables was only 2, then the equivalence limit for the Hotelling’s \(T^2\) test can be interpreted as the number of standard deviations. Hence, half a standard deviation is a relatively small amount but not that small if the data is normally distributed.

2. Data Generation

Step 1: Generate data from 2 Multivariate normal distributions with 8 means and pooled sample covariance matrix with the settings specified.

Step 2: Perform the IUT and Max Test, and Hotelling’s \(T^2\) test, for the Max Test we use
the algorithm given below:

*step 1:* Generate \( n \) samples of \( X_{1s}^*, \ldots, X_{ssX}^* \) and \( m \) samples of \( Y_{1s}^*, \ldots, Y_{ssY}^* \) from

\[
\frac{c_1}{2}, \ldots, \frac{c_1}{2}, \frac{c_1}{2}, \ldots, \frac{c_1}{2}
\]

and \( \frac{-c_1}{2}, \ldots, \frac{-c_1}{2}, \frac{-c_1}{2}, \ldots, \frac{-c_1}{2} \), recall \( S^* \) is the pooled sample covariance matrix estimated from the simulated data, and \( c_i \) = the equivalence limit (note that the position of \( c/2, -c/2 \) will correspond to the position where the largest variance is in our sample. This choice is the most conservative and will allow us to ensure control the type I error close to nominal.

*step 2:* Compute

\[
M^* = \max_{1 \leq i \leq \rho} \frac{\hat{X}_i^* - \hat{Y}_i^*}{\hat{\sigma}_{ii}^* + \hat{\sigma}_{ii}^*}
\]

with

\[
\hat{S}^* = \frac{1}{m+n-2} \left( \sum_{i=1}^m (X_{1s}^* - \overline{X}_{1s}^*, \ldots, X_{ssX}^* - \overline{X}_{ssX}^*) (X_{1s}^* - \overline{X}_{1s}^*, \ldots, X_{ssX}^* - \overline{X}_{ssX}^*)' + \sum_{i=1}^n (Y_{1s}^* - \overline{Y}_{1s}^*, \ldots, Y_{ssY}^* - \overline{Y}_{ssY}^*) (Y_{1s}^* - \overline{Y}_{1s}^*, \ldots, Y_{ssY}^* - \overline{Y}_{ssY}^*)' \right)
\]

so that \( \hat{S}_{ij}^* \) is the \( ij \)th element of \( \hat{S}^* \), and \( \hat{X}_i = m^{-1} \sum_{s=1}^m (X_{1s}^*, \ldots, X_{ssX}^*), \hat{Y} = n^{-1} \sum_{i=1}^n (Y_{1s}^*, \ldots, Y_{ssY}^*) \) are the sample mean vectors estimated from the bootstrap data.

*step 3:* Repeat steps 1-2 10,000 number of times.

*step 4:* Order the \( M^* \)'s from smallest to largest \( M^*(1) < M^*(2) < \ldots < M^*(10,000) \)
Step 5: \( M^*(\alpha(10,000)) \) is the sample estimate of \( M_a \)
RESULTS

To compare the 3 tests: the Intersection Union test, Max Test, and Hotelling’s $T^2$ test a simulation study was carried out using values for the parameters outlined in the simulation section.

Note that for the type I error estimation the equivalence limit for the data generation was set to be 40 for the Max Test and the IUT. The equivalence limit for Hotelling’s $T^2$ was varied so that the equivalence limit would be as close as possible to the estimated Mahalanobis Distance corresponding to a maximum mean difference of 40.

1. The simulated type I error is generally conservative for the IUT, on target for Hotelling’s $T^2$ test, and sometimes inflated for the Max Test.

Table 2 indicates that, in general the simulated type I error rates of the IUT are overly conservative. This is especially true of the IUT type I error results using the equal mean difference mean pattern. The Max Test has simulated type I error rates mostly on target but has slightly inflated rates when the variance is large and the mean pattern is either the real mean pattern or one mean pattern. Finally, the type I error of the Hotelling’s T test is very close to being on target regardless of the set of parameters or mean difference pattern.

2. Simulated Power of the Max Test and IUT is influenced by the pattern of the mean differences.

Simulated power is at its lowest values for the Max Test and IUT when the mean differences follow the same mean difference pattern, regardless of the variance-
covariance settings. On the other hand, power is at its highest values for the Max Test and IUT when the mean differences follow the one mean difference pattern. For example, power of the Max Test and the IUT increases by roughly 5-10% going from the same mean difference pattern (Figure 2) to real mean difference pattern (Figure 3) and increases by about 3-5% going from real mean difference pattern (Figure 3) to one mean difference pattern (Figure 4) regardless of the other parameter settings. The power of Hotelling’s $T^2$ test does not appear to change in any direction with regard to the mean difference pattern.

3. **Simulated Power increases for Hotelling’s $T^2$ test and decreases for Max Test and IUT as variance increases.**

Simulated Power increases for Hotelling’s test when the variance is increased regardless of the mean pattern generation. Figure 2, 3, and 4 illustrate this result. For example in Figure 3, power increases from ~10% to ~90% as the variance increases from 1k to 12k for a greatest mean value of 20, and real mean difference pattern. Conversely, power decreases for the Max Test and IUT as variance increases. For both of these tests this is true, regardless of the mean pattern, and the correlation (Figures 2, 3, 4). To illustrate, in Figure 3, power decreases from nearly 100% to about 10% for both the Max Test and the IUT as variance increases from 1k to 12k for a greatest mean value of 20.

4. **Simulated Power Increases for Max Test and IUT as correlation increases.**

The more highly correlated the data are the higher the power for the Max Test and IUT. The results are similar for all mean patterns (Figure 2, 3, 4), thus, as an example, consider Figure 3. Power of the Max Test increases from ~20% to ~60% when the
correlation increase from 0.05 to 0.7 for a variance of 6k, greatest mean difference of 20, and mean pattern real. Similarly, and for the same parameter settings, the power of the IUT increases from ~0 to ~20% as correlation goes from 0.05 to 0.7.

5. Max Test has higher simulated power than the IUT for all the simulation settings.

Regardless of the parameters values simulated, the Max Test outperforms the IUT in terms of power.
APPLICATION - DATA ANALYSIS

All three multivariate equivalence tests were applied to the final iNO dataset, which represented the 500 to 749g weight class, and consisted of the P/F ratios of two treatment groups, NO (inhaled nitric oxide), and CO (the placebo gas) measured over time. More than 35% of the data was missing for any given weight strata. Thus, the stratum we chose to analyze was the one with the highest number of subjects, stratum 1. Since the variance-covariance matrix indicated relatively high variability among the data points, we chose to use the higher end of the equivalence limit 20% of the baseline value for the Max Test and IUT. We set the equivalence limit to be 0.5 for Hotelling’s test, which is equivalent to \( \frac{1}{2} \) a standard deviation if we were to ignore the correlation in the formulation of the Mahalanobis distance (MHD).

For alpha=0.05, and the chosen equivalence limits, all 3 tests could not detect significant equivalence between the 2 groups over time.

- Hotelling’s Test: \( p=0.3714 \), effect size=0.508606
- Max Test: \( p=0.4016 \), effect size=38.65
- Intersection Union Test: \( p=0.4821 \), effect size=38.65

Finally, to smooth out some of the variability in the dataset, we tried a 3 window moving
average (Figure 5) on the 8 time points, which resulted in 6 time points to be tested. A three window average was used to reduce the 8 time points to 6 time points. This was done to smooth some of the variability in the dataset as well as reduce the data to a smaller subset (albeit in this case not much smaller) of data. Additionally, as the mean P/F ratio difference is largest at only 1 hour after treatment was given, it makes sense to combine the average at this time point with the other time points to get a more balanced distribution of the P/F ratio.

For alpha=0.05, and the chosen equivalence limits, the BM and IUT detected equivalence and Hotelling’s Test did not.

- Hotelling’s Test: p=0.25, effect size= 0.5013545
- Intersection Union Test: p=0.031, effect size= 16.43750
- Bootstrap Method: p<0.0001, effect size= 16.43750
Figure 2: Plot of the Moving Average of the P/F Ratio over Time. A 3 window average was used to form the moving averages for Strata 1. CO represents the placebo gas group, and NO represents the inhaled nitric oxide group.
DISCUSSION.

In this paper we have developed an alternative to the existing multivariate equivalence tests, Max Test. We examined the performance of this test as well as the IUT and Hotelling’s $T^2$ test in the context of clinical trial data. The simulated power of the Max Test was higher than the IUT for all the parameter settings examined. Hotelling’s $T^2$ test showed comparable power, but a direct comparison between the Hotelling’s $T^2$ and the IUT and Max Test is not available due to the different specifications of the null hypotheses of these tests. We also applied Max Test, the IUT, and Hotelling’s $T^2$ test to the clinical trial dataset involving the comparison of inhaled nitric oxide and a placebo gas treatment measured across 8 time points. While the tests were unable to detect equivalence this could be likely due to the high variability relative to the actual mean differences in this dataset.

1. Choice of Equivalence Limit for Hotellings $T^2$, Max Test, and the IUT.

In the clinical trial setting discussed in this paper iNO is already considered a safe and standard treatment for improving and relieving respiratory distress in neonates. Thus, using a larger equivalence limit of say 20% of the baseline P/F ratio value is reasonable given that iNO is already considered a viable treatment option, and the alternative is a placebo gas that has not been approved by the FDA. The choice of the equivalence limit for Hotelling’s $T^2$ test is based on normal theory and the case where the number of variables is two and no correlation exists. If we could derive a method of choosing the
equivalence limit based on a function of the number of variables and the variance-
covariance matrix, it would be helpful in understanding and interpreting equivalence in
terms of the Mahalanobis Distance, which is the sticking point of Hotelling’s $T^2$ test, the
interpretation of the rejection region of the null hypothesis as a meaningful equivalence
region.

2. Max Test is an improvement over the IUT in terms of simulated power, but has inflated
type I error when the variance is high.

For the parameter settings that were used for this study, the Max Test outperformed the
IUT in terms of simulated power. However, Max Test also has inflated type I error for
some of the larger variance settings, in particular the highest variance setting of 12k. The
variance and correlation settings based on the 25$^{th}$, 50$^{th}$, and 75$^{th}$ percentiles of the
variance and correlation estimated from the real data. In this study we used a simple
covariance structure where we assumed equal variance and equal correlation. We chose
this simple structure to be able to view how the increase in both variance and correlation
affects power. Besides varying the variance and correlation under the equal assumption,
we included a setting that used the variance-covariance matrix from the real dataset, and
for this setting the type I error of the Max Test is higher than the nominal 0.05.

3. Limitations: Parameter Settings

A limitation of this study is that there were a limited number of parameters that were
varied. The sample size was not varied and neither was the covariance structure, we
assumed equal variances and correlations, with the exception of the real data estimate of
the variance-covariance matrix. It would be interesting to see how the simulated power behaves for different variance-covariance structures. Additionally, it might shed some light on the reason why the type I errors for the Max Test appears to be inflated for the higher variance settings.
Figure 3: Simulation Study Rejection Rates for IUT, Max Test and HT-Test. Equal mean differences refers to how the means were simulated in simulation study, section 1. Colored lines are different combinations of variance and correlation. “Var” represents variance, and “Cor” represents correlation. All results are shown for alpha=0.05.
**Figure 4:** Simulation Study Rejection Rates for IUT, Max Test and HT-Test. Real mean differences refers to how the means were simulated in simulation study, section 1. Colored lines are different combinations of variance and correlation. “Var” represents variance, and “Cor” represents correlation. All results are shown for alpha=0.05.
Figure 5: Simulation Study Rejection Rates for IUT, Max Test and HT-Test. One mean differences refers to how the means were simulated in simulation study, section 1.

Colored lines are different combinations of variance and correlation. “Var” represents variance, and “Cor” represents correlation. All results are shown for alpha=0.05
<table>
<thead>
<tr>
<th>Max Test</th>
<th>variance, correlation</th>
<th>Equal</th>
<th>Real</th>
<th>One</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2=1000, \rho=0.05$</td>
<td>$&lt;0.001$</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=1000, \rho=0.40$</td>
<td>$&lt;0.001$</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=1000, \rho=0.70$</td>
<td>0.01</td>
<td>0.04</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=6000, \rho=0.05$</td>
<td>$&lt;0.001$</td>
<td>0.02</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=6000, \rho=0.40$</td>
<td>$&lt;0.001$</td>
<td>0.05</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=6000, \rho=0.70$</td>
<td>0.04</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=12000, \rho=0.05$</td>
<td>$&lt;0.001$</td>
<td>0.07</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=12000, \rho=0.40$</td>
<td>0.02</td>
<td>0.07</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2=12000, \rho=0.70$</td>
<td>0.03</td>
<td>0.08</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>var-covar data</td>
<td>0</td>
<td>0.07</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

| IUT | $\sigma^2=1000, \rho=0.05$ | $<0.001$ | 0.02 | 0.03 |
| $\sigma^2=1000, \rho=0.40$ | $<0.001$ | 0.02 | 0.04 |
| $\sigma^2=1000, \rho=0.70$ | 0.02 | 0.04 | 0.06 |
| $\sigma^2=6000, \rho=0.05$ | $<0.001$ | $<0.001$ | 0.01 |
| $\sigma^2=6000, \rho=0.40$ | $<0.001$ | $<0.001$ | 0.01 |
| $\sigma^2=6000, \rho=0.70$ | 0.02 | 0.01 | 0.01 |
| $\sigma^2=12000, \rho=0.05$ | $<0.001$ | $<0.001$ | $<0.001$ |
| $\sigma^2=12000, \rho=0.40$ | $<0.001$ | $<0.001$ | $<0.001$ |
| $\sigma^2=12000, \rho=0.70$ | $<0.001$ | $<0.001$ | $<0.001$ |
| var-covar data | 0 | 0 | 0.01 |

| Hotelling's $T^2$ | $\sigma^2=1000, \rho=0.05$ | 0.056 | 0.036 | 0.052 |
| $\sigma^2=1000, \rho=0.40$ | 0.04 | 0.054 | 0.052 |
| $\sigma^2=1000, \rho=0.70$ | 0.056 | 0.048 | 0.064 |
| $\sigma^2=6000, \rho=0.05$ | 0.048 | 0.056 | 0.06 |
| $\sigma^2=6000, \rho=0.40$ | 0.048 | 0.048 | 0.044 |
| $\sigma^2=6000, \rho=0.70$ | 0.056 | 0.024 | 0.04 |
| $\sigma^2=12000, \rho=0.05$ | 0.052 | 0.048 | 0.036 |
| $\sigma^2=12000, \rho=0.40$ | 0.058 | 0.04 | 0.062 |
| $\sigma^2=12000, \rho=0.70$ | 0.052 | 0.058 | 0.048 |
| var-covar data | 0.053 | 0.057 | 0.043 |

Table 2: Type I errors estimated from the simulation data. “Equal”, “Real”, and “One” refer to the mean pattern generation described in Simulation section 1. Alpha is set to 0.05. Type I errors were generated so that the data were as close to the boundary between the null and the alternative hypothesis as possible. The equivalence limit for the data generation was set to be 40 for the Max Test and the IUT and varied for the Hotellings T2 so that the equivalence limit would be as close to the estimated MHD as possible for each unique parameter setting.
References:


128


SUMMARY

This dissertation has followed the three-paper format. The goal of the first two papers is to better understand the behavior of equivalence tests, in particular multi-group equivalence tests, when applied to high-dimensional microarray data. In the first paper we derive a formula to relate the equivalence limits of both the F test and Range test to an already familiar and qualified measure of equivalence in the two group setting, the DR ratio. We also introduce an algorithm that employs Monte Carlo simulation to evaluate the Range statistic without having to make use of intensive computational methods such as numerical integration. Using both single gene and high dimension simulation study we found that the simulated power of the F test is higher than that of the Range test across all the parameter settings in our study; however, for a group size of 3 the F and Range test had very similar power for both simulation schemes. Both the F and Range test were incorporated into an R package, *EquivMulti* which allows users to perform the tests with the adaptations discussed in paper 1. The R package includes a sample low dimension dataset described in the book Testing Statistical Hypotheses of Equivalence and Noninferiority (Wellek 2010) as well as the sample microarray dataset, GSE1148 (Rinn et al. 2004) available on GEO. The functions and examples from the R package are summarized in paper 2. Finally, in paper 3 we explored multivariate equivalence tests. We developed a test, Max Test, to improve on the power of the IUT. It employs a parametric bootstrap to provide a better estimate the true distribution of the test statistic under the null hypothesis. We took this test and evaluated it along with the intersection union test and Hotelling’s $T^2$ using simulation study and real data analysis. At the outset of this dissertation work, we planned to apply and model our simulated data on
microarray data in a manner similar to what we did in paper 1. However, the multivariate
tests that we were working with would perform poorly, or not perform at all, on the
typical microarray data, which has high dimension data (a large number of variables) and
small sample sizes (which is a typical feature of microarray data). Thus, the application
data to which we applied our tests and modeled our simulation data from was a clinical
trial dataset involving the measure of oxygenation between a standard treatment, inhaled
nitric oxide (iNO), and a placebo gas (CO), over 8 time points (Cutter et al. 2006). The
simulated power of the Max Test was higher than that of the intersection union tests over
the range of the parameters modeled in our simulation study. On the other hand, for some
of the larger variance settings, Max Test has a somewhat inflated type I error (sometimes as
high as 0.08-0.09 depending on the seeds used to generate the data). Additional power
analysis was performed for Hotelling’s $T^2$ test, the results of the simulation study indicate
that Hotelling’s $T^2$ has decent power for the equivalence limit of 0.5 that was specified, and
type I error close to the nominal 0.05 level.

**Limitations**

In paper 1 we found that increasing group size increases the power of the F test. This
result seems to be a result of the equivalence criteria

$$e^2 = \frac{(k-1)}{2} \left( DR^2 - 1 \right)$$

we have derived for this test which is a function of group size. Further investigation into
this phenomenon might result in deriving an alternative form of the equivalence limit that
still depends on the group size but does not cause the test to be more liberal when group
size increases. Another limitation of paper 1 was the limited parameter settings
investigated. Since the data was simulated using real microarray data, the sample size, group size, and variance of the simulated data was fixed at the values observed in the real dataset. To further understand how these tests perform in the microarray setting additional representative microarray datasets should be used for the high dimension simulation. In particular, datasets that might have either high or low variability could be used to get a better handle on how variability across genes, sample size and group size in the microarray dataset affects the performance of these tests.

Paper 2 is an application paper for the R package, *EquivMulti*, which utilizes the functions created to do the F test and Range test of paper 1. Currently the package only computes the tests described in paper 1 with results output as a simple dataframe. Future iterations of the package might include additional output such as graphical summaries of equivalent genes for microarray dataset as well as confidence intervals for the effect sizes.

One of the main limitations of paper 3 was in defining a meaningful equivalence limit for Hotelling’s $T^2$ test. In our simulation study and analysis we used an equivalence limit of 0.5 with the reasoning that it is a reasonable limit to use when the interpretation of the effect size does not depend on the correlation (i.e. the correlation is zero) among the parameters. This reasoning is based on the understanding that without the correlation the effect size can be interpreted as the number of standard deviations the $\rho$ paired differences are from each other. However, this approach (not taking into account the correlation) is not taking advantage of the multivariate test. One way to derive an
equivalence limit that makes use of the correlation and the number of parameters both would be to keep a fixed equivalence limit of 0.5 say, and see how the simulated power behaves as the correlation structure, and number of parameters is varied. Then, if a pattern emerged as the correlation, variance, and group size increases or decreases, one might make a rule for the equivalence limit. Thus, if one wanted at least 80% power, one might increase or decrease the limit based on the rule for the equivalence limit as a function of the correlation matrix and the number of parameters. In our simulation study, we observed that as the variance increases the power of Hotelling’s $T^2$ test increases regardless of the other parameter settings. The pattern in power as a function of correlation or group size for Hotelling’s $T^2$ was not as clear for the parameter settings we used in our simulation study. Hence, further work might be done varying the variance-covariance structure and group sizes systematically to see if a pattern emerges for simulated power.

### Contributions

Contributions from the first paper include the adaptation of the difference ratio (DR) to the multigroup F and Range tests for equivalence, as well as exploring and measuring the performance of the tests in the microarray framework. The simulation study showed that a DR of 1.55 or higher will give moderate to high power for detecting relatively small values of means and mean difference among treatment groups. Additionally, it seems adequate power can be achieved with the Range test when group size is 3, but for group sizes greater
than 3 the F test is a more powerful test.

The second paper, the application paper, gives a description of the R package, *EquivMulti*, we created with the tests we explored in paper 1. The package gives investigators a way to easily implement the methods investigated in our first paper on both low and high dimensional data.

In the third paper we derived an alternative approach to the IUT multivariate hypothesis test. The test we developed, “Max Test” uses a parametric bootstrap that utilizes the estimated covariance matrix from the data. The type I error of the Max Test was slightly elevated for the high variance parameter settings; however, the simulated power of the Max Test was found to be superior to that of the IUT for all the parameter values modeled in our simulated datasets.

In conclusion, we have explored an area of hypothesis testing that is not widely known or employed. We’ve shed some light on how the multi-group equivalence tests perform on high dimensional data, and how the DR adapts to the multigroup setting in the context of gene expression analysis. We’ve created a fairly straightforward method for multivariate equivalence testing that improves on the intersection union test in terms of power, but should be further studied to determine what is the underlying cause of the inflated type I error at the higher variance settings.
DISCUSSION

The notion of equivalence itself is still a misunderstood concept. Papers are still concluding that there is “no difference” or that treatments are “the same” after a statistical hypothesis test for differences is found to be non significant. For example consider the topic of inhaled nitric oxide therapy in neonates that we discussed in paper 3 as an application for multivariate equivalence testing. A paper in the Journal of the American Medical Association compared mortality rates between patients receiving iNO treatment and a placebo using a standard statistical hypothesis test for differences. The outcome of the test was that no statistical difference could be detected for alpha set at 0.05. However, rather than stating that no statistical difference could be detected, the authors declared that mortality rates were “the same” between the two groups and cited a p-value of >0.05 as evidence of this “sameness” (Taylor R. W. et al. 2004). This is only one of many examples that can be found in the literature where a statistical test for difference has been misused for claiming equivalence. The misuse of the statistical hypothesis test for differences is often due to a lack of understanding of the hypothesis test itself. The hypothesis test was originally designed to control the number of false positives by essentially assuming the opposite of what is trying to be proved to be true (Neyman and Pearson 1933). Thus the typical test of difference can only allow two possible outcomes: rejection (at the specified alpha or significance value) or failure to reject. If the test results in the latter outcome, one may only determine that a significant difference was not detected by the test; there is no statistical justification for declaring there to be no difference. This is where a test of equivalence is an important and useful tool for a
variety of fields. A test of equivalence gives the statistical justification needed to assert that there is no difference among the treatments or conditions being compared. And indeed, more and more areas of research are utilizing equivalence tests. We have shown the application of equivalence in the field of genetic research using gene expression data and in clinical trials using treatment data measured over time; but these are just a couple of examples in a growing world of research and data collection where equivalence testing may be a valuable tool and method of determining similarity. For example in the field of public health policy, technology has enabled massive data collection. One of the main priorities of public health is to eliminate differences in immunization coverage among various populations regardless of their demographics. In order to show “practical equivalence” among various groups’ vaccination coverage Barker et al. (2004) used equivalence testing to show the differences between groups lied within some tolerable limit. A future application for equivalence testing in the field of public health research may be 30-day readmission rates.

Thirty-day readmission rates in hospitals are likely to become a prevalent topic in the quality of medical care literature for 2013. This interest may be due to the Affordable Health Care act. This act will make hospitals pay an increasing penalty of 1% to 3% per-year of their total cost for 30-day readmission medicare patients. Thus in the first year of 2012, roughly 2,200 hospital facilities were hit with a 1% penalty of their total Medicare costs related to patients that had 30-day readmissions. In 2013, the penalty will increase by 1%, up to a maximum of 3% per year. Healthcare researchers are developing scoring systems such as the IMRS-HF to help predict 30-day readmission chances. While this is
just one scoring system to help predict readmissions there are other systems attempting to predict 30-day readmission chances such as CORE’s 30-day cardiovascular readmission score (Rapp et al. 2011). Many scoring systems based on different models and this abundance of systems can lead to confusion for health analyst, physicians, and financial administrations due to the many models available. Statistical equivalence testing may be applied here to show test the hypothesis that the two or more scoring systems are equivalent. For example, we may be interested in predicting 30-day readmission chances for a group of older woman ages 60-65, having hypertension, and type II diabetes. Readmission chances for each of the women would be computed per model, and then a multivariate test of equivalence would be applied to show that their readmission chances are equivalent regardless of the model used. This would then let the health researchers save time by not having to choose between various models that give equivalent results.

While we have discussed the utility of equivalence testing, a practical examination of experimental design issues in equivalence focused clinical trials may help prevent invalid conclusions. Understanding experimental design issues in the area of clinical trials which can sabotage the validity of equivalence testing are important to principal investigators. For example, in non-inferiority trials, which are often thought of as a form of “one-sided” equivalence testing, one is commonly interested in testing if one standard treatment method is equivalent to a newer method of treatment. In the article “Pros and Cons of Non-Inferiority Trials” (Pocock 2004), Pocock goes over potential factors in non-inferiority trials that can invalidate conclusions obtained from an equivalence test. The first harmful factor to account for is to make sure that the selection of patients used for
the study were the same type of population for which the old treatment was demonstrated to be effective. If a newer population were to be used which hadn’t a demonstrated superior effect of the older treatment, then lack of equivalence may be due to differences in population rather than the results of the equivalence test. Another factor Pocock suggests is treatment compliance. In this situation the older treatment should be given to the patients in the same way it was administered before. If there are deviations from the way the older treatment was administered, then any conclusions from the equivalence tests could be a result of this discrepancy. There are other factors mentioned such as “outcome measures”, and “duration of treatment and evaluations.” In each these factors, the guiding principal is to make sure the methods used to establish the old treatment are consistently practiced and applied when assessing non-inferiority with the new method. Otherwise it’s hard to distinguish if true lack of equivalence comes from the test, or by differences in which the old treatment was implemented or measured during the experiment.

As a final point, one of the most challenging aspects of equivalence testing lies in the definition of the equivalence limit. In paper 1 we derived our equivalence limit using the definition of the difference ratio for population equivalence and in paper 3 we related the equivalence limit to a clinically significant value for the P/F ratio. Thus the definition of the equivalence limit is dependent upon both knowledge of the data and the type of hypotheses being tested. While the equivalence limit is challenging to define, it is also a valuable way to define how close treatments or outcomes are to each other. The original hypotheses tests were not created with a very holistic view of analyzing data in that they
are geared toward showing differences or change against a null hypothesis of no difference when often times showing treatments or conditions are within a range of values can be a very valuable and informative to investigators. Thus, perhaps what needs to be addressed is not the type of hypothesis test being performed but the manner in which the null hypothesis is stated. Perhaps the scientific community needs to be performing test where the testing is being done against an interval null hypothesis that would give more information on the range of values for which sameness or difference is achieved.
GENERAL REFERENCES


Michael T. Rapp, Jennifer A. Mattera and Sharon-Lise T. Normand


Schuirmann, D. J. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* 37:617.


