Spring 2014

A Comparison of Two Statistical Tests for Interaction in Genetic Data

Clair Smith
Western Washington University

A Comparison of Two Statistical Tests for Interaction in Genetic Data

Clair Smith

Dr. Amy Anderson

June 6, 2014

Introduction

This paper focuses on statistical methods that test for the effect of a single gene in a way that accounts for interaction with other genes. Such tests of association can be difficult since there may be many genetic and environmental factors that contribute to an effect. A gene is a hereditary DNA sequence that codes for a specific protein. A locus is a gene's location in the DNA sequence of nucleotides (A, T, G, and C) and an allele is a specific version of a gene that has multiple forms. The existence of interactions between loci makes it difficult to determine the effect of a single genetic factor. Gene-gene interaction occurs when two genes together affect how a trait is expressed. That is, their combined effect is not the same as the sum of the two genes' individual effects. If there are multiple genes that effect a genetic trait, then examining a gene in isolation without allowing for interactions may result in missing the gene's effect. This is why allowing for interaction in a test of association can be more successful than tests that do not allow for interaction.

The genetic effect we are interested in is a person's response to a medical treatment. We have reason to believe that a gene-gene interaction affects whether or not an individual responds to treatment. In order to test this, we use statistics to determine if a certain combination of genes is associated with response to treatment. The goal of such a statistical test is to be able to say "people with this combination of genes are X times more likely to respond to treatment." Rather than genes, we will be looking for interaction between single-nucleotide polymorphisms (SNPs). A SNP is a certain sequence of DNA that has a nucleotide that varies among individuals. For example, some individuals might have the sequence AACGC at a certain locus while others may have the sequence AACCC. We say that this SNP has two alleles. Most SNPs have only two alleles. We wrote a program that performs two different statistical tests of associaion between predetermined SNPs of interest and treatment outcome.

Regression is a method that is commonly used to determine association between an outcome variable such as response to treatment and predictor variables such as SNPs. We use regression to fit genetic data using some model (equation) relating the outcome to the predictors. For our purposes, the outcome, Y, will be whether or not an individual responds to a treatment. If individual i responds to the treatment then $Y_i = 1$ and if they do not respond to the treatment then $Y_i = 0$. The predictor variables, Z and W, are SNPs and take on values of 0, 1, or 2. We call the first SNP, Z, the "primary" SNP and the second SNP, W, the "modifying" SNP. Because the response veriable is binary, we use logistic regression to fit the data.

The probability that individual i responds to the treatment, $P(Y_i = 1)$, is given by a function of Z_i and W_i. The probablity model is

$$P(Y_i = 1) = \frac{exp(\alpha + \beta_1 Z_i + \beta_2 W_i + \beta_3 Z_i W_i)}{1 + exp(\alpha + \beta_1 Z_i + \beta_2 W_i + \beta_3 Z_i W_i)}, \tag{1}$$

where exp is the exponential function, α is a constant, β_1 and β_2 are the coefficients of the predictors W and Z, and β_3 is the coefficient of the interaction term ZW. Including the interaction term allows us to detect whether gene-gene interaction between the SNPs contributes to association with treatment outcome. There are two different tests we could use to detect such associations. Testing for interaction means determining whether or not there exists a statistical interaction between the two predictor variables. In the above model, this would correspond to testing whether the coefficient of the interaction term, β_3, is zero. Conversely, testing for association in the presence of an interaction means testing simultaneously for a main effect and an interaction rather than for an interaction alone. This corresponds to testing whether β_1 and β_3 are both

zero. To do this, we compare the fit to the data of the model in which the main effects of the two SNPs and their interaction are included with a model in which all the terms involving the primary SNP are removed, i.e. $\beta_1 = \beta_3 = 0$. If this latter model fits the data better than the previous model, then we say that the primary SNP has no effect on whether or not an individual responds to treatment.

Chapman and Clayton [2007] discuss a different method for association tests with multiple SNPs. Their method develops a "pseudo-score" test statistic that tests for association between the primary SNP and response to treatment as well as for an interaction between the primary and a modifying SNP. The primary SNP is considered to be associated with response to treatment if the primary SNP itself is associated with response or if the interaction of the primary SNP with a modifying SNP is associated with response. This differs from other methods that simply test whether or not an interaction between the two SNPs has an effect on response to treatment. Chapman and Clayton tested the interaction term only among individuals who responded to treatment. They called this the "case-only test for interaction." They were able to do this with the assumption that the primary and modifying SNPs are independent. To adjust for the multiple tests for the different SNPs, they take the maximum of the psuedo-score tests across all modifying SNPs as the overall test of effect for the primary SNP.

We compare the Chapman and Clayton model for the direct case to the logistic probability model given by (1). The comparison is based on the power of the two tests. The power of a statistical test is the probability that it detects an association between the outcome and the predictors when an association actually exists in nature. In a sense, the power is a measure of how effectively you are using the data. Higher power means there is a better chance of detecting the set of SNPs that are associated with response to treatment. Chapman and Clayton claimed that the case-only test for interaction increases the power of the test since fewer individuals are considered, and therefore, the variance is decreased. The powers of the tests we are comparing are the probabilities that they detect an association between response to treatment and the SNPs when there truly is an association present in the data.

Methods

Creating a Random Dataset

We had the program generate random data sets so that we could determine the powers of the two tests. The program allows us to choose whether or not there is an association between the SNPs and response to treatment. When generating a dataset, the program uses a random number generator and the probabilities of having the different versions of the two SNPs to determine an individual's Z and W values. Each SNP has two alleles, say "a" and "A." Since we inherit a set of DNA from both our mother and our father, there are three possible genotypes. An individual could have the "aA" genotype, the "aa" genotype, or the "AA" genotype. For each SNP, we coded the "aa" genotype as a 2, the "aA" genotype as a 1, and the "AA" genotype as a 0. For both the primary and the modifying SNP, we set the probability of being an "aa" genotype to 0.25 and the probability of being an "aA" genotype to 0.75. First, the program sets every individual's genotype to "AA", that is, every individual had a value of 0 for both the primary and modifying SNP. To determine an individual's genotype at one of the SNPs, a random number between 0 and 1 is generated. If this number is less than 0.25, then the value for this individual's genotype is assigned a 1. Otherwise, if the number is greater than 0.25 but less than 0.75, then the value of the individual's genotype is assigned a 2. If the random number is greater than 0.75, then the value of the individual's genotype remains a 0.

To determine an individual's Y-value, the program uses the logistic probability model (given above) with the individual's Z and W values to calculate the probability that the individual responds to treatment. The program then produces a random number between 0 and 1 and if this number is less than the probability that the individual responds to treatment calculated in the previous step, then their Y-value is assigned a 1.

In order to compare the powers of the Chapman and Clayton method and the logistic regression method, our program generates random datasets then analyses the datasets using the two different tests. We set the values for the parameters α, β_1, β_2 and β_3 along with the probabilities of having the different genotypes of each of the SNPs and then create a random dataset based off of these values. The resulting datasets include 250 individuals along with their Y-value ($Y = 1$ for response and $Y = 0$ for no response) as well as which versions of the primary SNP and modifying SNP that they have. The next step is to analyse the datasets

to determine the power of the two different methods.

Analyzing the Data

Once a dataset is generated, the program calculates a test statistic from the Chapman and Clayton psuedo-score test and a test statistic from the logistic regression. A test statistic is a numerical summary of a data set. The logistic regression uses a Chi-squared test statistic which we get from comparing the probability model given in (1) with the model containing only the W term, (the modifying SNP). The Chi-squared test is a test of whether the full model fits the data better than the smaller model. A p-value is then calculated from this test. The p-value of a statistical test is the probability that we would see a test statistic at least as extreme as the one observed given there is no association between the SNPs and treatment outcome in the data. We call this situation in which ther is no association the "null hypothesis". A small p-value (less than 0.05) indicates that there is a very small chance that we would see a test statistic as extreme as the one observed if there was no association. We call the probability 0.05 the "statistical significance" of the test. A p-value less than 0.05 gives strong evidence that there truly is an association between having a certain combination of SNPs and responding to treatment in the data. In other words, it is very unlikely that we would observe such an extreme test statistic simply due to chance if there truly was no association in the data. A p-value less than 0.05 is considered to be significant evidence to reject the null hypothesis. Thus, if there is less than a 5% chance we would see a test statistic at least as extreme under the null hypothesis, then it is likely that there is an association in the data.

In order to test whether the primary SNP is associated with treatment response, the direct pseudo-score test determines whether or not β_1 and β_3 are both zero. If they were zero, then the probability that an individual responds to treatment would be the same regardless of which version of the primary SNP they have. Thus, there would be no association between the primary SNP and treatment outcome. There are two components to Chapman and Clayton's pseudo-score. The first component is a score for testing whether or not β_1 is zero, and the second component is a score for testing whether or not β_3 is zero. Testing whether or not β_3 is zero is the same as testing whether or not an interaction between the two SNPs contributes to the probability that an individual responds to treatment. If there is more than one modifying SNP, then the maximum of the psuedo-scores is taken as the overall test for the effect of the primary SNP. Chapman and Clayton were able to determine an equation for the psuedo-score and the variances of its two components.

After obtaining the test statistics and p-values from the logistic regression and the psuedo-score test for the dataset, the program performs a permutation test. Permutation testing allows us to get a p-value for a test without knowing how the data is distributed. Our program allows us to choose whether or not there is an association in the data. This allows us to make an association in the data then see how well the two tests detect the association. Given a data set, the program first permutes (scrambles) the individual's Z values. By permuting the Z values, we are forcing there to be no association between SNP genotype and treatment outcome since the Z values have been reassigned. Thus, we are forcing the null hypothesis to be true.

Next, test statistics and p-values are obtained from the two different tests. The Z values are then permuted once again and new p-values are calculated. This process is carried out 1000 times. We can calculate the number of times that the p-values for the two tests were less than or equal to 0.05, that is, the number of times the null hypothesis was rejected. The number of times that the null hypothesis was rejected out of 1000 permutations is the overall p-value for a test obtained through permutation testing. Since we make the null hypothesis of no association true by permuting the Z values, it would make sense that we would reject the null hypothsis a very small number of times due to chance if our test is working well. That is, there would be a very small number of instances where we would see data that gives a test statistic as extreme the one observed for the original data set before permuting.

Since we have control over the values of the β's, we can give them positive values to make the null hypothesis of no association false. For example, having a postive value for β_1 would cause people with larger values of Z, such as 1 or 2, to have greater odds of responding to treatment than people with a Z-value of zero. Therefore, it would be true that genotype at the primary SNP is associated with treatment outcome in this scenario. The probabilty that we reject the null hypothesis in such a situation where association

exists is the power of a test. That is, the power of a test is the probability that we see data that allows us to reject the null hypothesis of no association when it is false. We calculate the power of the two different tests by counting the number of times permutation testing resulted in a p-value less than or equal to 0.05, the significance level, out of 2000 simulations of the program. For each of the 2000 simulations, p-values are obtained for the two tests from a permutation test with 1000 iterations. Plotting these powers for different values of a parameter results in a power curve for each test. If one test has a power curve that sits above the other when we plot power against values of β_1, then that test is better at detecting association when the null hypothesis is false (when there is association).

Results

In order to determine which test is more likely to detect an association, we compared the power of the psuedo-score test to that of the logistic regression for values of β_1 between 0 and 1. Figure 1 shows the power curves of the two tests holding the other parameters constant. We hold β_2 and β_3 at zero and α at -2. The graph shows that we have higher power to detect the association betwen response to treatment and the primary SNP with the psuedo-score test. For extreme values of β_1 near zero and one, the two tests have approximately equal powers. The figure shows that when the null hypothesis is true, that is, when $\beta_1 = \beta_2 = 0$, both tests have the same significance level. So the proportion of times that the null hypothesis is rejected is 0.05, the significance level of the tests.

Figure 2 shows the power curves of the two tests for values of β_3, the interaction term coefficient, between 0 and 0.7. We hold α at -2 and β_1 and β_2 at 0. The psuedo-score test has only slightly better power when detecting the association between response to treatment and an interaction between the primary and modifying SNPs. At $\beta_3 = 0.4$ the psuedo-score test has a significantly greater power than the logistic regression (p-value$= 0.0017$). But when $\beta_3 = 0.5$, the two tests do not have significantly different powers (p-value$= 0.056$).

Figure 3 shows that the psuedo-score test has lower power than the logistic regression for values of β_3 greater than 0.4 when we make $\beta_2 = 2$ and $\beta_1 = 0$. The powers of the two tests are lower when we make $\beta_2 = 2$ since the interaction term is no longer the only term contributing to a response to treatment. Having a positive coefficient for the modifying SNP creates a more complicated relationship between genotype and response to treatment. In the logisitic regreassion, we compare the model with all of the terms to the model with only the modifying SNP term. If the modifying SNP has a positive effect on response to treatment, then the model with only this term will fit the data well. This results in a smaller p-value since the p-value is the probability that we reject the null hypothesis that the full model fits better than the model with only the modifying SNP term. A smaller p-value means higher power. The logistic regression model had higher power than the psuedo-score test because the logistic regression compares the fit of the full model to the model with only the modifying SNP. The psuedo-score test is set up to test for an effect at the primary locus while allowing for interaction and since the primary locus had no effect ($\beta_1 = 0$), then the psuedo-score test had lower power than the logistic regression in this case.

Conclusion

Because there may be many genetic factors that contribute to an effect, finding a single gene that is associated with the effect can be difficult. The existence of gene-gene interactions may cause researchers to miss the effect of a genetic factor if they do not account for interaction in their test. Testing for association while allowing for interaction is a way to discover such gene-gene interactions that contribute to an effect such as response to treatment. Chapman and Clayton's psuedo-score method is better at detecting association in the presence of an interaction than the logistic regression method. This is true because the pseudo-score test had higher power than the logistic regression test in most situations. The psuedo-score test had lower power when there was no effect at the primary SNP, but a positive effect for the interaction and modifying SNP terms, that is when $\beta_1 = 0$ shown in Figure 3. If an interaction term is contributing to an effect then it is likely that both the SNPs in the interaction contribute individually to the effect. This is why we include both the primary and modifying SNP terms, as well as their interaction in the logistic regression model. It

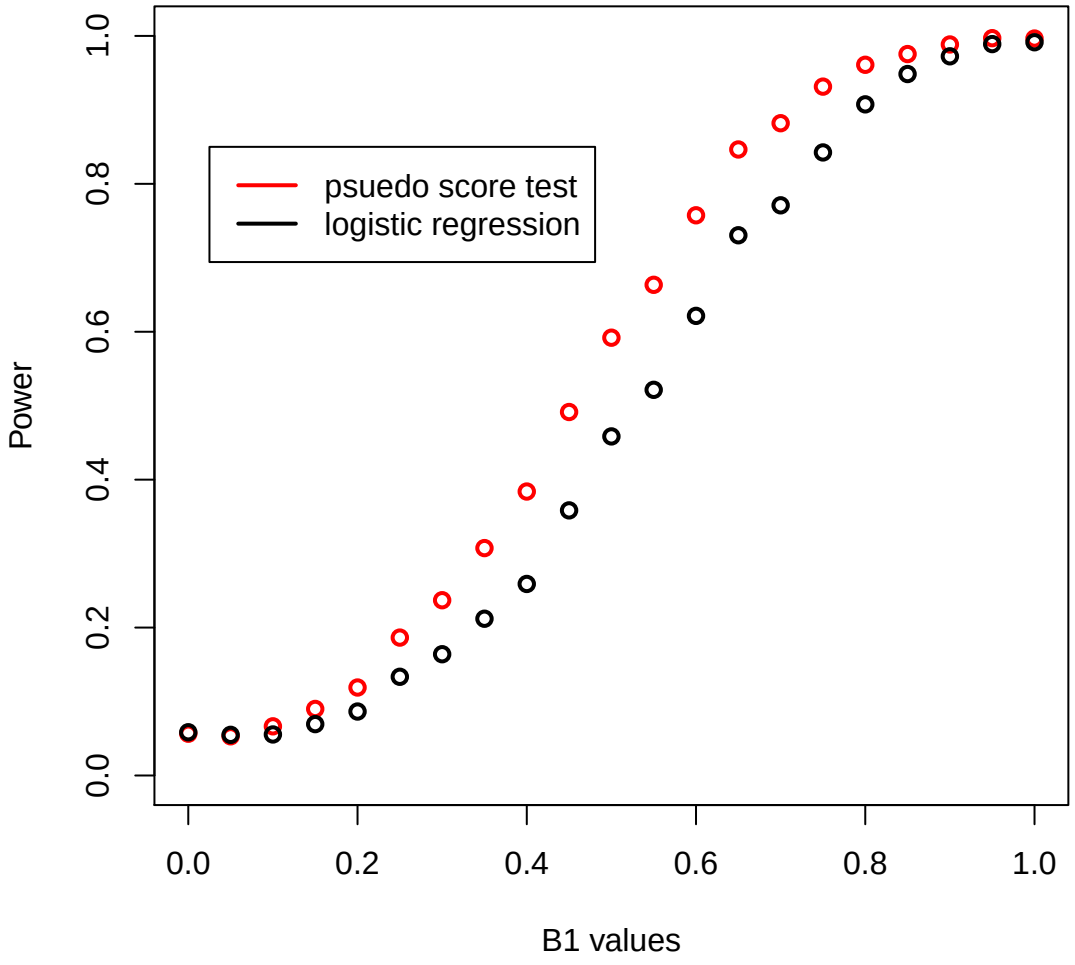

Figure 1: Power curve for values of β_1 for $\beta_3 = \beta_2 = 0$ and $\alpha = -2$.

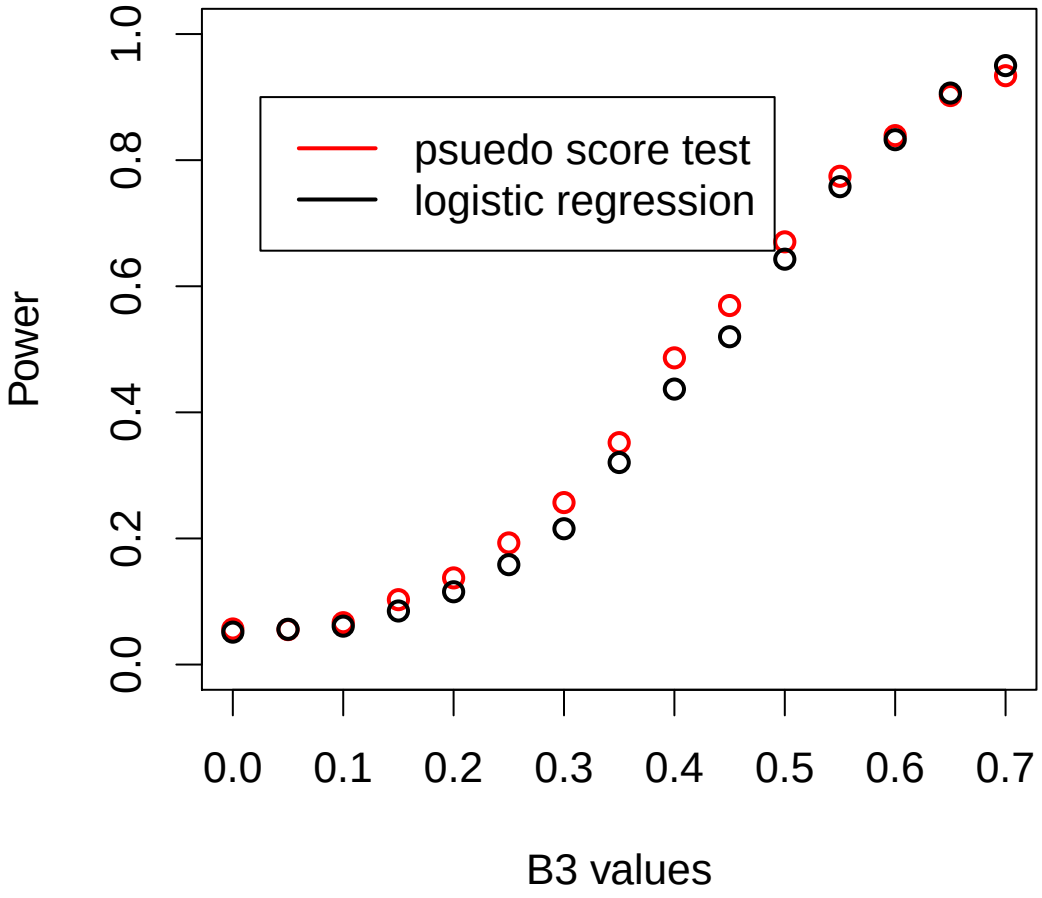

Figure 2: Power curve for values of β_3 for $\beta_1 = \beta_2 = 0$ and $\alpha = -2$.

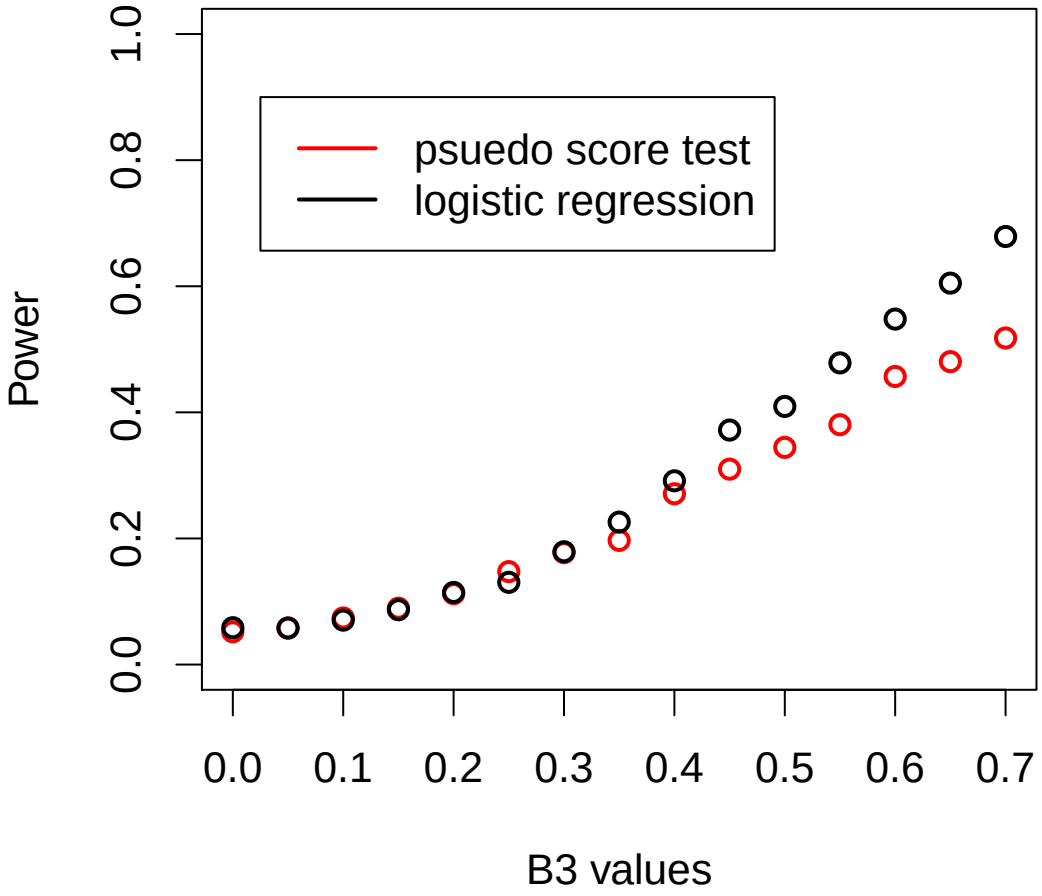

Figure 3: Power curve for values of β_3 for $\beta_1 = 0$, $\beta_2 = 2$ and $\alpha = -2$.

is unlikely to see a situation in which the interaction contributes to the effect but one of the SNPs has no individual effect. Thus, the psuedo-score test is generally better at finding association in genetic data when an interaction is present.

Discussion

Chapman and Clayton's psuedo-score test is designed to search for pairwise interactions that may contribute to an effect from a set of many SNPs. Their test is set up so that there is a primary SNP that is thought to contribute to an effect, and a pool of modifying SNPs that may interact with the primary SNP. The psuedo-score test takes the maximum score out of all the modifying SNPs as the overall test for an effect at the primary locus. We considered a situation in which there was only one modifying SNP, but given more time, we could make power curves for situations in which there are many modifying SNPs. In this situation, the program would take the maximum of the psuedo-score tests and the maximum of the logistic regression tests as the overall tests of the primary SNP.

This paper only considered one particular model of gene-gene interaction, and we generated our data based upon this same model, see (1). It would be interesting to see if our results change when the true model of gene-gene interaction differs from the model simulated. For example, we could add another modifying SNP term and more interaction terms between the three SNPs.

Given more time, we would use the program with real medical data to test for association while allowing for interaction. This would require coding the data so that it matches the program and adjusting the program so that it imports the medical data.

Despite these considerations, the results shown indicate that the Chapman and Clayton psuedo-score test for association in the presence of interaction is better than the logistic regression method in most cases. Such tests for gene-gene interaction may be useful for detecting loci involved in complex disease.

References

Chapman, Juliet, and David Clayton. "Detecting Association Using Epistatic Information." *Genetic Epidemiology* 31.8 (2007): 894-909.