DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Statistical Methods for Analyzing DNA Methylation Data and Subpopulation Analysis of Continuous, Binary and Count Data for Clinical Trials

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation	Yip, Wai-Ki. 2015. Statistical Methods for Analyzing DNA Methylation Data and Subpopulation Analysis of Continuous, Binary and Count Data for Clinical Trials. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:14226106
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Statistical Methods for Analyzing DNA Methylation Data and Subpopulation Analysis of Continuous, Binary and Count Data for Clinical Trials

A dissertation presented

by

Wai-Ki Yip

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

December 4, 2014

©2014 - Wai-Ki Yip
All rights reserved.

Dissertation Advisor: Professor Christoph Lange Wai-Ki Yip

Statistical Methods for Analyzing DNA Methylation Data and Subpopulation Analysis of Continuous, Binary and Count Data for Clinical Trials

Abstract

DNA methylation may represent an important contributor to the missing heritability described in complex trait genetics. However, technology to measure DNA methylation has outpaced statistical methods for analysis. Novel methodologies are required to accommodate this growing volume of DNA methylation data. In this dissertation, I propose two novel methods to analyze DNA methylation data: (1) a new statistic based on spatial location information of DNA methylation sites to detect differentially methylated regions in the genome in case and control studies; and (2) a principal component approach for the detection of unknown substructure in DNA methylation data. For each method, I review existing ones and demonstrate the efficacy of my proposed method using simulation and data application.

Medical research is increasingly focused on personalizing the care of patients. A better understanding of the interaction between treatment and patient specific prognostic factors will enable practitioners to expand the availability of tailored therapies improving patient outcomes. The Subpopulation Treatment Effect Pattern Plot (STEPP) approach was developed to allow researchers to investigate the heterogeneity of treatment effects on survival outcomes across increasing values of a continuously measured covariate, such as biomarker measurement. I extend the STEPP approach to continuous, binary and count outcomes which can be easily modeled with generalized linear models (GLM). The sta-

tistical significance of any observed heterogeneity of treatment effect is assessed using permutation tests. The method is implemented in the R software package (**stepp**) and is available in R version 3.1.1. The efficacy of my STEPP extension is demonstrated by using simulation and data application.

Contents

List of Figures

List of Tables

In memory of Yip Peng-Kai and Mak Shau-Lan.
Without their dediction to my education,
none of this would have been possible.

Acknowledgments

I would like to first thank my wife, Elaine, and my children, Christine and Charles, for their love and support over the years that I was pursuing my Ph.D. degree at Harvard. I'd like especially to thank my advisor, Prof. Christoph Lange, who has served as a wonderful mentor over the years and without whom none of this would be possible. Next, I'd like to thank my dissertation committee members, Prof. Nan Laird and Prof. Rich Gelber to serve on my committee, and continue to be my mentors, collaborators, and friends. I'd like to give special thanks to Jelena Tillotson-Follweiler, Florence Yong, Mark Meyer, Dandi Qiao, Yared Gurmu, Pita-Juarez, Christina McIntosh, Emmanuel Dimont, Ann Lazar, Xin Victoria Wang, William Barcella, Dawn Demeo, Martin Aryee, Bernard Cole, Marco Bonetti and David Wypij who all have provided an abundance of advice and moral support over the past few years. Finally, I'd like to thank my friends and extended family who have given me encouragement over these years.

1. A Novel Method for Detecting Association Between DNA Methylation and Diseases Using Spatial Information

Wai-Ki Yip[1], Heide Fier[2], Dawn L. DeMeo[3], Martin Aryee[4], Nan Laird[1]and Christoph Lange[1]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, United States of America
[2]Department of Genomic Mathematics, University of Bonn, Bonn, Germany
[3]Channing Division of Network Medicine, Brigham and Women Hospital, Boston, MA, United States of America
[4]Massachusetts General Hospital, Boston, MA, United States of America

Abstract

DNA methylation may represent an important contributor to the missing heritability described in complex trait genetics. However, technology to measure DNA methylation has outpaced statistical methods for analysis. Taking advantage of the recent finding that methylated sites cluster together, we propose a Spatial Clustering Method (SCM) to detect differentially methylated regions (DMRs) in the genome in case and control studies using spatial location information. This new method compares the distribution of distances in cases and controls between DNA methylation marks in the genomic region of interest. A statistic is computed based on these distances. Proper type I error rate is maintained and statistical significance is evaluated using permutation test. The effectiveness of the proposed SCM is evaluated by a simulation study. By simulating a simple disease model, we demonstrate that SCM has good power to detect DMRs associated with the disease. Finally, we applied the SCM to an exploratory analysis of chromosome 14 from a colorectal cancer data set and identified statistically significant genomic regions. Identification of these regions should lead to a better understanding of methylated sites and their contribution to disease. The SCM can be used as a reliable statistical method for the identification of DMRs associated with disease states in exploratory epigenetic analyses.

1.1 Introduction

For the past decade, the Genome-Wide Association Test (GWAS) for population-based genetic anlaysis has identified many genetic factors associated with complex diseases (Manolio, Brooks and Collins, 2008). However, DNA polymorphisms have explained only a portion of inheritance patterns in many of these diseases, leaving a large amount of heritability still to be accounted for (Mahar, 2008). Some of the missing heritability might be explained by epigenetic mechanisms, such as DNA methylation, that control gene expression by means other than changes to the DNA sequence (Manolio et al.,

2009). Most human DNA methylation marks are likely stable, making them useful as disease biomarkers. Analyzing data from cancer patients, we observe a widespread disruption of the human DNA methylation profile. The human DNA methylation profile plays a role in the etiology of numerous other complex diseases including asthma, coronary heart disease, and bipolar disorder (Foley et al., 2009). However, there is currently no "gold-standard" statistical approach to identify human DNA methylation profiles that are associated with diseases. The need for better analytical methods for studying DNA methylation has grown, as recent advances have made large-scale measurements much more tractable.

Recent advances in biomedical technology make it possible to perform large-scale measurements of DNA methylation across the human genome. As a result, the methylated state of each CpG site and its location in the human genome can be identified (Laird, 2010). Thus, some of the methods used to detect disease susceptibility loci for genetic variants can be applied to detect disease susceptibility CpG sites. However, there are numerous challenges in analyzing DNA methylation data: measurement errors and batch effects are common for DNA methylation data, methylation effects may not be captured completely if surrogate tissue samples are used, and contamination from cellular heterogeneity in samples has led to spurious conclusions. Furthermore, methylation levels are clustered (Hackenberg et al., 2010). If we can take advantage of distance information, we can devise a more powerful test to detect association.

Instead of trying to identify CpG sites individually, we propose screening the genome for potential regions worthy of detailed analysis by utilizing spatial location information of CpG sites. Distances between CpG sites, measured in DNA base pairs, provide us with crucial information about where DNA methylation events occur. There are two reasons why this information is important. First, CpG sites tend to cluster together in promoter regions that affect gene expression. Second, studies have shown gene expression levels are associated with genomic regions of variable methylation instead of a single CpG site. For example, the level of methylation in the promoter region of a gene may be inversely

associated with the level of gene expression of that gene (Eckhardt et al., 2006; Hansen et al., 2011; Lister et al., 2009). A comprehensive framework for a DNA methylation analysis pipeline is proposed by Jaffe (Jaffe et al., 2012). In their approach, the user specifies a statistical model. Our method is similar to Jaffe's method in that we are "bump hunting"' along the genome with the goal of detecting differentially methylated regions (DMRs). The main difference is that we are using specific spatial information to come up with a robust statistic.

Our proposed approach is adapted from an algorithm which successfully detects rare genetic variants associated with diseases using spatial information (Fier et al., 2012). The algorithm is based on genetic distance between polymorphic single nucleotide polymorphism (SNP) variants and their distribution. We believe the distributions of genetic distance between CpG sites are similar. In epidemiology, identification and quantification of patterns in disease occurrence provide the first steps toward increased understanding of that particular disease. As stated, "The location where an event happens may provide some indication as to why that particular event occurs. Thus spatial statistical methods offer a means for researchers to use locational information to detect and quantify patterns in public health data, and to investigate the degree of association between potential risk factors and diseases" (Waller and Gotway, 2004). Spatial location information for CpG sites may be the additional data needed for a robust statistic for DNA methylation.

We call our method the Spatial Clustering Method (SCM). By creating a statistic that incorporates spatial location information and methylation measurement at each CpG site, our method can identify DMRs in the genome that are candidates of association with diseases. Since the location is measured from the beginning of each chromosome, the analysis is restricted to one chromosome at a time. In the following sections, we describe the SCM in detail and highlight its potential. We evaluate the SCM with a simulation study demonstrating that it has good power to detect DMRs while maintaining the prespecified type I error rate. To test our approach, we apply the SCM to a publicly available clinical data set with colorectal cancer as the phenotype. In doing this, we identify significant

DMRs of the human genome potentially linked to this disease. By reliably and specifically identifying methylated regions associated with disease, rather than individual sites, SCM may facilitate the study of causal associations between methylation and disease.

1.2 Methods

In this section, we will discuss the conceptual development of the method and the detailed algorithm to compute the SCM statistic.

1.2.1 Conceptual Development

By observing patterns of genetic events in our genome, we can identify and quantify patterns in disease occurrence. We adapt some of the statistical methods developed for spatial analysis to help our investigation of association between potential genetics factors and diseases. There are two categories of data in spatial data analysis: feature data consisting of geographical information that uniquely identifies the location where events occur, and attribute data consisting of counts of measured information about the events. To frame genetic investigation into the spatial data analysis framework, we need a "map" with relevant feature information including location and distance. Then, genetic events can be identified together with the location of occurrence and related with one another based on topological constraints.

In our model, CpG methylation marks can be viewed as spatial information arranged as points along the chromosome when it is stretched into a line. Thus, an analysis that explicitly uses spatial information can be very informative. In particular, we wish to detect whether the set of locations observed contains clusters of events reflecting areas with associated increases in the likelihood of occurrence (e.g., unusual aggregations of cases of a particular disease).

Array-based platforms to assay DNA methylation result in a percent methylation at a single CpG site; for our method we need to transform these percent methylation values to methylation units. The transformation is done by using a weight for each CpG site, developed based on the percent methylation values of controls and the site's distance from the nearest neighboring CpG site. The weight is then applied to each CpG site transforming the percent methylation value to methylation units for that site regardless of whether it is a case or control. Each methylation unit is counted as an occurrence of an event.

With the basic entities defined, the next step is to characterize location patterns. The traditional spatial descriptions of the location patterns such as the intensity functions or the K-functions are based on geographical map information and may not be suitable for genetic analysis. So, we adapted an idea that has been applied successfully for the development of a nonparametric association test for genetic rare variants to our method (Fier et al., 2012). The location pattern is characterized by the distance distribution between each consecutive event on the line. A distance vector is a vector of distance between events from all the methylation measurements in the genome region of interest. In our current adapatation of the rare variant method, we create two distance vectors - one is for methylation units from the cases and one for methylation units from controls. These two distance vectors represent samples from the distance distribution for events in case and control groups.

With case-control point data, we can compare the control locations which provide baseline information on spatial patterns of the population at risk and the case locations which provide spatial patterns of the disease. As done in many other epidemiology studies, an association test that identifies clusters of CpG sites and examines their association with diseases can now be constructed based on differences between the control and the case distance distributions. The idea is to develop a test statistic that captures the differences between these two distance distribution functions. If the differences are statistically significant, we can reject the null hypothesis that distributions are the same, and conclude

that the genomic region is associated with the disease phenotype.

With the distance vectors developed, we can use the Ansari-Bradley statistic to test if there is a difference in dispersion (Ansari and Bradley, 1960). Since the choice of creating the weight based on the control group is arbitrary, we repeat the same procedure using a weighting scheme based on percent methylation values and nearest neighboring CpG site from cases. The final statistics is computed as the maximum of the two derived statistics. Since we assumed that the CpG sites are clustered with the same association effect in the same genomic region, the differences between the distribution functions are magnified for the small genomic distances. Therefore, the test captures information for both distances and methylation association effects.

As in most other spatial tests, the distribution of the difference in distribution functions does not have a closed analytic form. Thus, the statistical significance of the proposed test statistic is obtained by Monte Carlo/permutation method that randomly assigns case/control status to the study population while maintaining the total number of cases and controls and the DNA methylation structure of the genomic region of interest.

For exploratory analysis, we can set up a genomic window which contains a fixed number of CpG sites. Then, we scan each of the chromosomes from the beginning to the end using a sliding genomic window to look for regions that are significantly different for further analyses. However, it is important to apply multiple testing corrections for this approach [Kuan et al 2012].

1.2.2 Detailed Algorithm

In order to explain our idea fully, we present the detailed algorithm using the following notation: We assume that a defined genomic region has been assayed or sequenced in N subjects in the context of a case-control study, recording the physical positions in DNA base-pairs from the beginning of each chromosome of all the CpG sites in the

region, and the methylation levels at each site for all subjects. Assume that there are a total of K CpG sites under consideration. The sorted position vector for the CpG sites is denoted by $P = (p_1, \ldots, p_i, \ldots, p_K)$ and index by $1 \leq i \leq K$. The subjects in the study are indexed by $1 \leq j \leq N$. To identify cases and controls, we define the indicator functions $I_{case}(s_j) = 1$ if the subject s_j is from the case group and $= 0$ otherwise, and $I_{control}(s_j) = 1$ if the subject s_j is from the control group and $= 0$ otherwise. The methylation signal vector corresponding to the position vector, P, for subject, j, and is denoted by $M_j = (m_{1,j}, \ldots, m_{i,j}, \ldots, m_{K,j})$.

Computing the test statistic

First, we developed a weight for each CpG site based on information obtained from the control group. We denoted the weight based on the control for the i^{th} CpG site by $w_{0,i}$. The derivation of the formula for the weights is explained later in the section. By applying this weight to each CpG site, we transform the percent methylation value for cases and control into methylation units for the i^{th} CpG site denoted by $u_{case\ or\ control,i}$ given by taking the largest integer from the following formula:

$$u_{case,i} = w_{0,i} * \sum_j (m_{i,j} * I_{case}(s_j)),$$

and

$$u_{control,i} = w_{0,i} * \sum_j (m_{i,j} * I_{control}(s_j)) \tag{1.1}$$

Since each methylation unit is considered as an occurrence of an event, we have an ordering of events based on the position order of the CpG sites. Next, we create a sequence of distances, du_{case}, between each event for cases and a sequence of distances, $du_{control}$, between each event for controls. Each of these two sequences of distances can be treated

as samples from the corresponding distance distribution.

$$du_{case} = \{\ldots,\ distance\ in\ base-pair\ between\ the\ i^{th}\ event\ and\ i+1^{th}\ event\ for\ cases, \ldots\},$$

and

$$du_{control} = \{\ldots,\ distance\ in\ base-pair\ between\ the\ i^{th}\ event\ and\ i+1^{th}\ event\ for\ control, \ldots\} \tag{1.2}$$

The applied weighting schemes only influence the skewness of the derived distance distribution functions, and have no impact on the values of the observed non-zero distances. The null and alternative hypotheses are:

$$H_0 : the\ distribution\ functions\ of\ cases\ and\ controls\ are\ the\ same$$
$$(or\ no\ association\ of\ methylation\ pattern\ and\ disease)$$
$$H_A : they\ are\ different$$

To test our hypothesis, we used the Ansari-Bradley test which is a non-parametric, two-sample test on the variability of the distance distribution functions. If the distribution functions were different, we expect to see one of the samples to be more spread-out than the other. We applied the test to our two samples, du_{case} and $du_{control}$. Since the samples are obtained using a weighing scheme based on subjects from the control group, we denoted the resulting Ansari-Bradley statistics by $A_{control}$. The weighting scheme described above was based on subjects from the control group. We also derived the weight based on subjects from the case group. Using the same notation, we obtained the weight based on the cases for the i^{th} CpG site by $w_{1,i}$ and transformed the percent methylation values to methylation units for the i^{th} CpG site denoted by v_i :

$$v_{case,i} = w_{1,i} * \sum_j (m_{i,j} * I_{case}(s_j)),$$

and

$$v_{control,i} = w_{1,i} * \sum_{j} (m_{i,j} * I_{control}(s_j)) \tag{1.3}$$

With the implicit ordering of events based on CpG site positions, we created a sequence of distances, dv_{case} , between each event for cases and a sequence of distances, $dv_{control}$, between each event for controls.

$$dv_{case} = \{\ldots,\ distance\ in\ base-pair\ between\ the\ i^{th}\ event\ and\ i+1^{th}\ event\ for\ cases, \ldots\},$$

and

$$dv_{control} = \{\ldots,\ distance\ in\ base-pair\ between\ the\ i^{th}\ event\ and\ i+1^{th}\ event\ for\ control, \ldots\} \tag{1.4}$$

Similarly, we applied the Ansari-Bradley test and obtain the statistics, A_{cases}. This tests for the differences between the distance distribution functions for cases and controls when weighted by the subjects from cases. The final statistic is the maximum of the two:

$$A_{final} = max(A_{cases}, A_{control})$$

A rejection of the null hypothesis implies an association of the tested methylation sites in the region with affection status. Because of the likely presence of tied observations in the distance distribution functions, we used an implementation of the standard Streitberg/Roehmel shift algorithm to retrieve exact distribution functions for our test statistics (Streitberg and Roehmel, 1986). To obtain the significance of the test statistics, we used permutation testing. For each specified genomic region, case/control status was randomly assigned to each individual so that the total number of cases and controls of the original study is maintained. As a result, the methylation structure of the study-population samples was kept constant. The p-value was estimated as the proportion of

permutation test statistics which were more extreme than the actual observed test statistic for the data.

Computing the weight

In order to incorporate the importance of spatial proximity between CpG sites and to control for an uneven distribution of the percent methylation values, we defined weights that depend on both cluster and percent methylation values. We took the distance to the nearest neighbor for each of the CpG sites and the shortest distance vector, $D = \{d_i\}$ for all $2 \leq i \leq K-1$ where

$$\begin{aligned} d_i &= min(|p_i - p_{i1}|, |p_i - p_{i+1}|) \\ d_1 &= |p_2 - p_1| \\ d_K &= |p_K - p_{K-1}| \end{aligned}$$

Then, we computed the mean percent methylation value, ms_{case}, of all subjects from the case group and mean percent methylation value, $ms_{control}$, of subjects from the control group separately:

$$ms_{case} = \sum_i \sum_j (m_{i,j} * I_{case}(s_j)) / \sum_i \sum_j (I_{case}(s_j)),$$

and

$$ms_{control} = \sum_i \sum_j (m_{i,j} * I_{control}(s_j)) / \sum_i \sum_j (I_{control}(s_j)) \tag{1.5}$$

We standardized each methylation sites percent methylation value with respect to the mean of the cases. Our weighting scheme for the ith CpG site based on the subjects from cases, was given by

$$w_{1,i} = 1 + ([ms_{case}] / [\sum_j (m_{i,j} * I_{case}(s_j) + 1])) / log(d_i + 1),$$

and based on the subjects from controls, was given by

$$w_{0,i} = 1 + ([ms_{control}]/[\sum_{j} (m_{i,j} * I_{control}(s_j) + 1])/log(d_i + 1) \quad (1.6)$$

The weights are created to more highly weight close by CpG sites with lower methylation values.

1.2.3 Simulation Study

In order to evaluate the proposed SCM, we simulated the null population where the disease phenotype was independent of the percent methylation value at each CpG site. Then, we simulated the disease populations where certain sites were selected to associate with the disease status. The proposed test statistic was computed on 1000 simulated samples so that we could evaluate both the type I error rate and examine the power of the test to detect associations with the disease phenotype.

For exploratory purposes, we varied the size of the genomic region of interest centered on the chosen CpG sites in our simulation study. For most of our reported results, we used a genomic region of interest of 81 consecutive sites, denoted by K, centered with the chosen CpG site. For this simulation study we used the es_ICGN dataset, which contains 325 controls and 620 cases with chronic obstructive lung disease (COPD). The es_ICGN data set contains data collected for the International COPD Genetics Network with 1085 Caucasian subjects using the Illumina infinium27K beadchip (Qiu et al., 2012). Es_ICGN is a family-based study of subjects with ages between 45 and 65 with at least 5 pack-years of cigarette smoking. In our analysis, we treated all subjects in the dataset as independent cases and all unaffected siblings as controls. This should have negligible impact on our simulation study. The dataset has 26486 CpG sites after data cleaning and a total of 1085 subjects. We only used data from the 945 subjects with clean data for our analysis.

Simulation of the null population

Currently, there is no standard paradigm to simulate methylation marks in the human

genome. In order to simulate the null population, we used the following approach. A random small genomic region might provide the null region (region that is not associated with the phenotype) that we were looking for. Using the methylation measurements for a specific CpG site for all subjects, we approximated the distribution of the methylation percent values for that site. We picked three CpG sites from chromosomes from the beginning, middle and the end of the human genome (the 1300^{th} CpG site of chromosome 1, the 200^{th} CpG site of chromosome 14, and the 800^{th} CpG site of chromosome 19) to be the center of genomic region of interest for our study. We simulated samples using a normal distribution $N(\mu, \sigma^2)$ with μ and σ^2 estimated from the data and then truncated any values outside of 0 and 1. Thus, the active range is much smaller than [0, 1]. The es_ICGN dataset provided the genomic positional information and the empirical distribution of the percent methylation values of all the CpG sites. Using this approach, we generated a methylation percent value for each subject. This was applied to all the sites in the genomic region that we were interested in. Since the CpG sites are relatively sparse in this dataset, we treated sites within subjects as independent. Since the null model stipulates that there was no effect of methylation on the disease outcome, in order to achieve maximum power, we randomly allocated the 945 subjects to controls (473) and cases (472). We generated 1000 replicates of each null sample.

Simulation of the disease population

There is also no standard approach on how to simulate methylation marks on the DNA molecule that are associated with the disease. To evaluate our test statistics, we chose a simple disease model. By assuming that the mean of the percent methylation values of CpG sites selected as associated with the disease are shifted by the same constant percentage of their standard deviation from the null, we simulated methylation data for case subjects based on the null distribution with the mean shifted upward. Since the SCM detects the difference in methylation level, this is the same as shifting the mean upward for control subjects.

Following the same convention established in the null population simulation, we considered a genomic region with K number of CpG sites without knowing in advance how many of the sites are associated with the disease and their effects. Thus, the number of disease-related CpG sites and their effects are being varied in this simulation and are controlled by the following parameters: D denotes the number of disease-related CpG sites in the genomic region of interest while S denotes the percentage of standard deviation (SD) shift in the mean of cases from the null distribution (assumed to be the same for all disease-related CpG sites).

To simplify the simulation, we assumed the disease-related CpG sites were adjacent to one another and at the center of the genomic region of interest while the rest of the CpG sites under investigation were not associated with the disease. This is designed for maximum power. Future research can be done to test against other alternatives with disease-related CpG sites not centered or not adjacent to one another. We first generated the null percent methylation values using null empirical distributions based on the estimates from the es_ICGN dataset for all K CpG sites. To obtain our case sample, we replaced the percent methylation value for the selected disease site for each subject from cases by percent methylation value generated using a distribution that was based on the null empirical distribution but with a varying shift in the mean which was S % of the standard deviation (SD). We repeated the same procedure for all the other disease-related CpG sites, where the percent methylation values at each site were independently selected. Similar to our null population simulation, we generated 1000 replicates of each case sample.

1.2.4 Application to a colorectal cancer dataset

To illustrate the feasibility of the SCM to real data, we applied it to an exploratory analysis of a colorectal data set, which is publicly available from the Cancer Genome Atlas (TCGA) website (TCGA Network, 2012). The data was processed using Illumina Infinium 450K chip. There are a total of 329 samples from cancer patients. To avoid subject heterogeneity,

we used only 76 matched samples from the same patients one sample from solid tumor and a corresponding sample from normal tissue. For simplicity and illustrative purposes, we treated them as independent samples. The SCM can be extended to handle matched pair but is beyond the scope of this study. These samples were processed as 3 separate batches. We used the ComBat function in the R package sva (Leek et al., 2012) to correct for batch effect. To illustrate our method, we applied it to all the methylation marks (methylome) on chromosome 14 in our data set using a fixed sliding window approach. Each sliding window, representing a genomic region of interest, contained W CpG sites. The first window of chromosome 14 started at the 1^{st} CpG site and runs through the Wth CpG site; the second window of chromosome 14 started at the 2nd CpG site and runs through the $W + 1^{th}$ site; and so on till the end of the chromosome.

1.3 Results

1.3.1 Evaluation of type I error under the null

We applied the SCM to all the simulated null samples. Using an α-level of 0.05, the test maintained a type I error rate of about 5%. Using the notation K for the total number of consecutive CpG sites included in the genomic region of interest, we experimented with varying the values of K from 3 to 81 in our studies. The type I error rates were 0.052, 0.048 and 0.054 for the three chosen regions with $K = 81$. Similar results for type I error rate were obtained for values of K as small as 3. Based on the simulation results, we concluded that the SCM captured the type I error rate properly at the 0.05 significance level regardless of the size of selected region.

1.3.2 Power Estimates

To estimate power, we used the sites around the 1300th CpG site of chromosome 1. We chose K to be 51 and let D, the number of disease related CpG sites, vary from 1 to 21.

Table 1.1: Power estimates for detecting disease sites within 51 CpG sites around chr 1, 1300th CpG site

S	$D = 1$	$D = 3$	$D = 5$	$D = 11$	$D = 21$
0%	0.052	0.052	0.048	0.051	0.057
10%	0.075	0.096	0.231	0.459	0.854
20%	0.087	0.155	0.608	0.912	1.000
30%	0.124	0.266	0.891	1.000	1.000
50%	0.214	0.512	0.999	1.000	1.000
100%	0.537	0.951	1.000	1.000	1.000

To be precise, with $K = 51$ and $D = 1$, there are 25 null sites followed by 1 disease-related site, the chosen site, and followed by 25 more null sites. We shifted the mean of the percent DNA methylation value as a percentage (0%, 10%, 20%, 30%, 50% and 100%) of the standard deviation.

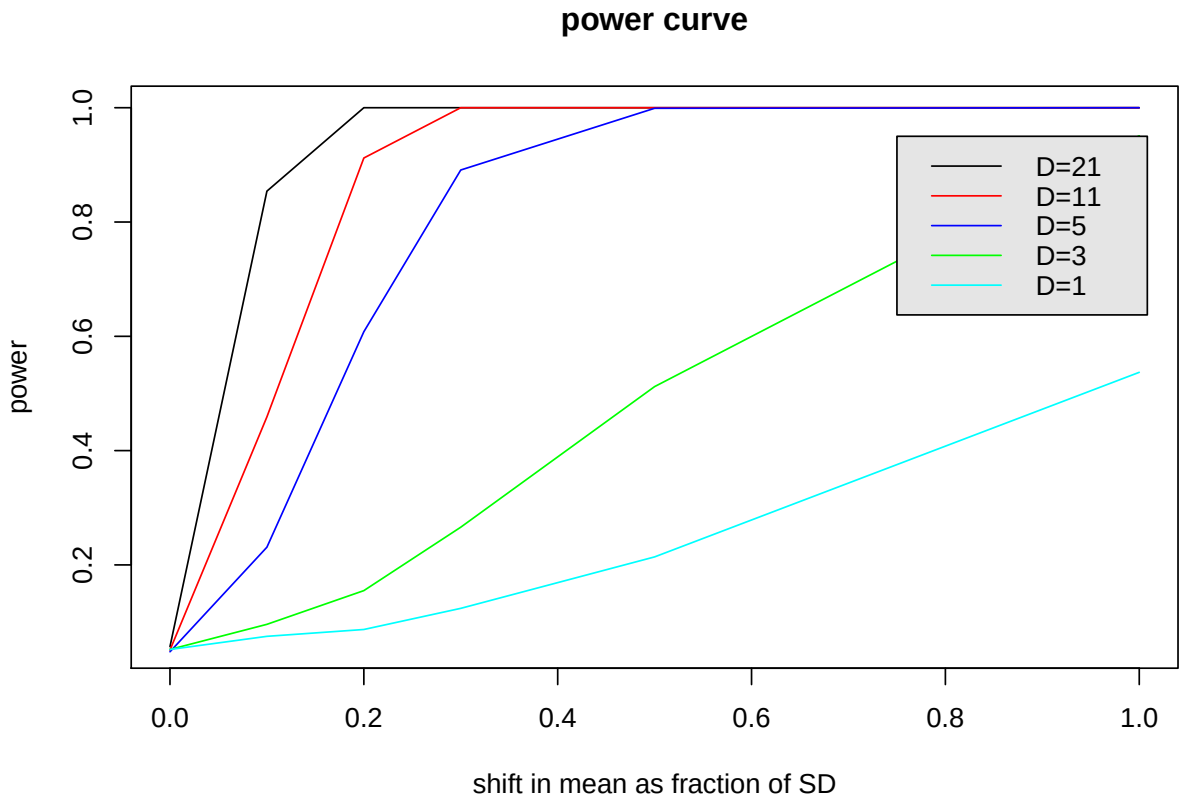

Figure 1.1: Power curve as the mean of percent methylation value is shifted. The color indicates the number of disease sites within the window (of 51 CpG sites) of investigation.

If there is only 1 disease CpG site among the 51 sites that we were investigating, the power to detect the association is very small. As expected, this parameter combination did not differ much from the null. However, as the number of disease sites increases, the power to detect the association increased significantly even if the amount of shift in the mean methylation value was small. Note that this was a simple model that assumed the sites were sampled independently and no correlation structure of the disease associated sites was taken into account. So, for complex diseases with a few disease sites (<5) and a small difference in percent methylation values ($< 50\%$ from the standard deviation), the power was low (i.e. less than 80%). We demonstrated that this method is good at detecting association if there are a number of disease CpG sites (> 5) clustered together affecting

the outcome of the disease. The details were shown in table 1.1 and figure 1.1.

1.3.3 Application of the SCM to chromosome 14 of a cancer dataset

For illustrative purposes, we applied the SCM to chromosome 14 of a colorectal cancer dataset from TCGA (TCGA Network, 2012). The study characterizes somatic alterations in colorectal carcinoma and identified 32 somatic recurrently mutated genes. For methylation patterns, the paper reported the identification of 4 subgroups based on unsupervised clustering of the promoter DNA methylation profiles of 236 colorectal tumors but not direct association. Since DNA methylation profiles were disrupted extensively by cancer, we expected our method to show diverse number of DMRs. As we have demonstrated our result for null sample work with values of K ranging between 3 sites and 81 sites, we chose the number of consecutive CpG sites, K, as 51, somewhere in the middle of our studied range. We scanned the entire chromosome 14 from beginning to the end.

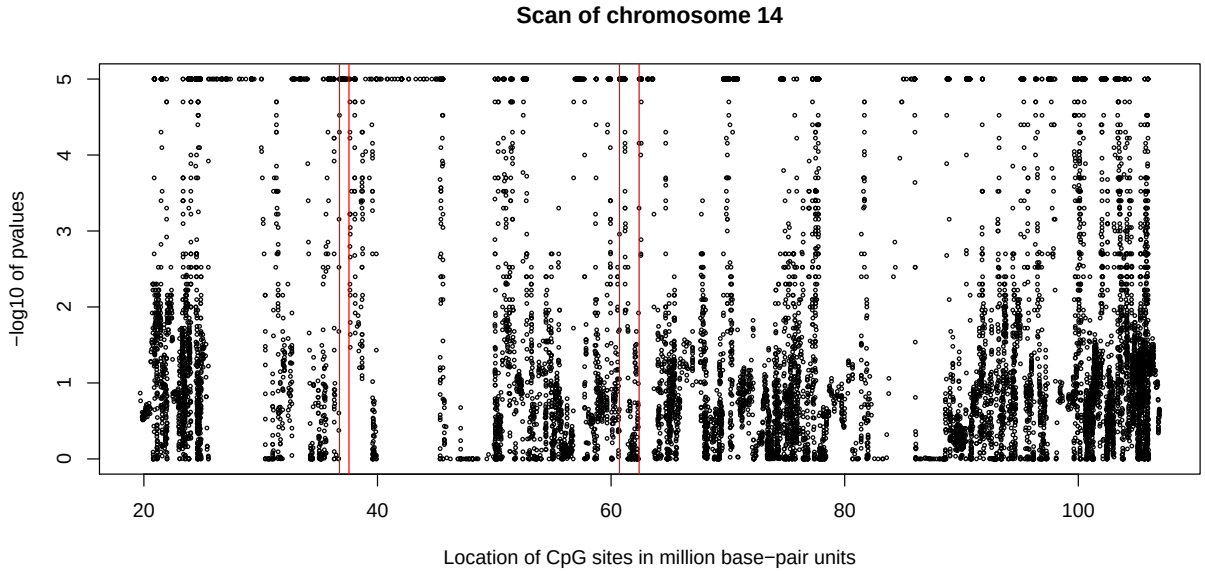

Figure 1.2: Using windows of 51 CpG sites, the sliding window scan shows regions of chromosome 14 which have p-values of $< 10^{-5}$ by using the SCM.

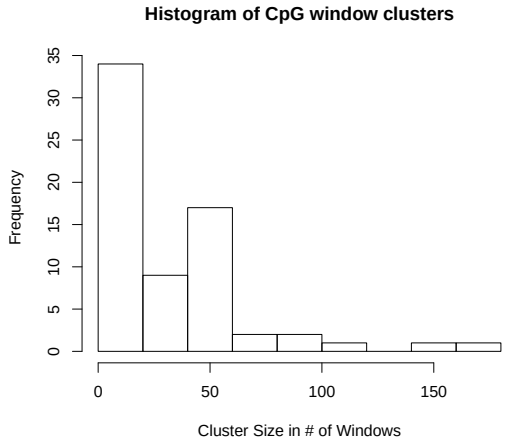

Figure 1.3: Histogram of the size of CpG window clusters on chromosome 14 with p-values $< 10^{-5}$.

The scanning result showed that there were 2079 windows with pvalues $< 10^{-5}$, which is the Bonferroni corrected significance level. Many of these significant windows were contiguous forming 67 clusters of statistically significant regions. Two windows belong to a window cluster if they have 4 or less non-significant windows in between. The two largest window clusters were comprised of 176 and 144 windows. These two window clusters were shown as the red vertical lines in figure 1.2. The distribution of the window cluster sizes was shown in figure 1.3.

This exploratory analysis showed that there were two big window clusters with sizes > 140 CpG sites on chromosome 14 which were associated with the disease. To show direct association, detailed modelling and analysis work should be done on these two particular DMRs. One window here consists of 51 consecutive CpG sites which usually cover several genes. For example, the biggest window cluster of DMRs that was detected on chromosome 14 started at location 22958402 and ended at 23398747 on chromosome 14 with a width of 440345 base-pair units in the region of $14q11.2$. It comprises of the following 10 genes: AJUBA, PSMB5, ACIN1, CEBPE, SLC7AB, BCL2L2, PABPN1, IL25, CMTM5, and MYH6. So, the SCM used in this exploratory investigation helped to identify genomic regions that were associated with colorectal cancer.

A similar analysis is done on chromosome 10 and the results are available in the Appendix A.1.1.

1.4 Discussion

It is well known that human DNA methylation CpG sites are spatially clustered (Hansen et al., 2011; Lister et al., 2009). Thus, spatial location information of the CpG sites should help to identify the associations between disease and differentially methylated regions. Our proposed Spatial Clustering Method, SCM, combines the use of physical spatial location information and percent methylation values from CpG sites of a genomic region to create a statistic. This statistic is then used to assess the significance of the relationship. By incorporating spatial location information, the SCM improves the likelihood of finding real significant associations. As shown in the simulation result, the SCM has reasonable power to detect methylated regions which had differences in the mean level of percent methylation values.

The SCM is a simple and easy to use statistical test for comparing genomic regions. No modeling is needed. As demonstrated in the simulation study, the SCM maintains proper type I error rate. Compared with a simple GWAS-approach which treats all CpG sites as independent and a linear mixed model with estimated correlation, one may end up with an inflated type I error. And if the disease associated CpG sites are spatially clustered and the shift in mean percent methylation values is associated with the disease, the SCM method will have good power to detect the association. The SCM can also be used for screening association across the entire genome. By applying a sliding window approach, we can locate genomic regions of high significance quickly. In the colorectal cancer data set, we used the SCM to identify two regions with large clusters in chromosome 14 of a colorectal data set. Once the SCM identifies DMRs, rigorous follow up such as biological analysis on the identified regions can be used to find possible causative agents for disease. Our simulation is based on a simple shift in the mean percent methylation values and if

the disease process causes such events to occur, the SCM will have good power to detect these DMRs. But, one must be cautious to interpret the p-values from a genomic region scan since a single CpG site can belong to a number of these sliding windows. Multiple testing adjustments must be applied. Thus, using the SCM for screening is suitable at this point as an exploratory tool and identifies DMRs for further investigation using additional biological methods. Further research should be done to examine the power of SCM in other measures besides a simple shift in the mean.

Although the SCM does not allow the user to adjust for batch effects directly, this can be corrected by using a number of standard software packages such as sva (Leek et al., 2012) to the data before applying the SCM. In order to avoid confounding, one can match samples by the confounding covariate or by a propensity score. However, if the experimental design is not under your control, one can adjust for potentially confounding covariates, such as gender, by applying less powerful statistical technique such as stratification. The p-value is computed using the permutation approach. It requires a significant amount of computing time. On an Intel i7 2.90GHz CPU, it has taken about 40 hours to complete a scan on the entire chromosome 14 with 10000 permutation using a sliding window size of 51 CpG sites on the TCGA data set. So, this is not yet practical to do a genome-wide scan for sliding windows sequentially. However, since the computations are independent of one another, one can break up the computations into parallel processes to get the result faster. The actual performance depends on the computing platform, processors available, amount of memory available, and the precision one would like to obtain for the p-value. Furthermore, we expect to continue fine-tuning the algorithm by taking advantage of parallelization and improving efficiency. These preprocessing works are common for statistical analysis and should not incur a big burden on the user. Before this can be used as a general association test, we need to determine how population and family substructure affect DNA methylation profiles in future studies. If we ignore these effects, it can lead to spurious conclusions. Furthermore, since DNA methylation profiles can change over time, we cannot conclude definitively that DNA methylation are in the causal pathway to

the disease. The reverse could also be true - the disease could have caused the variability in DNA methylation profile. Thus, for now, the SCM is best used as an exploratory screening method to locate potential associated DMRs for further analysis.

However, there are some limitations imposed by the current technology for providing full information on the spatial distribution of CpG locations as only predefined CpG sites are on the chip. The SCM is best suited for high coverage chip or sequencing data where more location information of CpG sites is available. For example, data from the denser Illumina Infinium 450K chip provides more information than data from a sparser Illumina Infinium 27K chip. Additional extensions to the basic algorithm can be made to further enhance the SCMs power. For example, different kinds of window patterns can be used for screening analysis, in addition to the sliding window pattern. The advantages and disadvantages of using different kinds of window patterns will need to be studied in a future research. Another extension is to study when it is appropriate to treat small percent methylation values as totally unmethylated. The state of methylation at a particular CpG site is binary in nature, just on or off, but the percent methylation value for a CpG site is continuous. Due to the uncertainties associated with the signal, a small value is reported even if the methylated state is off. Thus, small percent methylation values could be an artifact of the measuring process. The distribution of distances to neighboring CpG sites will change if there are more zero value methylation sites. The power to detect association may increase if we can treat these CpG sites with very small percent methylation values as totally unmethylated. By expanding our approach, we can further increase the power and utilities of the SCM.

By incorporating spatial location information into the analysis, we create a novel and simple association test that locate biologically relevant DNA methylation segments in the genomic region of interest. It captures type 1 error rate properly and has power to detect if there is a shift in the mean methylation value in the genomic region of interest. Locating the DMR is the essential step that can lead to the discovery of the underlying biological process between DNA methylation and diseases.

Acknowledgements

We would like to thank Dr Emily Wan for her assistance in the preparation of the COPD data set and the TCGA Research Network for making the COAD dataset publicly available. Generation of the methylation data included in the simulation (es_ICGN) was supported by NIH/NHLBI R01 HL089438. DLD is additionally supported by P01 HL 105339 and P01 HL 083069. We would also like to acknowledge the support of Wai-Ki Yip by the Clinical Epidemiology of Lung Diseases Training Grant (T32 HL 007427). International COPD Genetics Network (ICGN) investigators: Edwin K. Silverman, Brigham & Women's Hospital, Boston, MA, USA; David A. Lomas, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK; Barry J. Make, National Jewish Medical and Research Center, Denver, CO, USA; Alvar Agusti and Jaume Sauleda, Hospital Universitari Son Dureta, Fundacin Caubet-Cimera and Ciber Enfermedades Respiratorias, Spain; Peter M.A. Calverley, University of Liverpool, UK; Claudio F. Donner, Division of Pulmonary Disease, S. Maugeri Foundation, Veruno (NO), Italy; Robert D. Levy, University of British Columbia, Vancouver, Canada; Peter D. Par, University of British Columbia, Vancouver, Canada; Stephen Rennard, Section of Pulmonary & Critical Care, University of Nebraska Medical Center, Omaha, NE, USA; Jrgen Vestbo, Department of Cardiology and Respiratory Medicine, Hvidovre Hospital, Copenhagen, Denmark; Emiel F.M. Wouters, University Hospital Maastricht, The Netherlands.

References

Ansari, A., and Bradley, R. (1960). Rank-sum tests for dispersions. *Am Math Statist* **31:**1174–1189.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA and others, P. (2006). DNA methylation profiling of human chromosomes 6, 20, and 22. *Nat Genet* **38:**1378–1385.

Foley DL, Craig JM, Morley R, Olsson CJ, Dwyer T, Smith K and Saffery R (2009). Prospects for epigenetic epidemiology. *Am J Epidemiol* **169:**389–400. doi:10.1093/aje/kwn380

Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, Silverman EK, Pagano M, Mangold E, Lange C (2012). 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft plate. *Bioinformatics*, **28(23):**3027–3033.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H Liu Y, Diep D and others (2011). Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43:**768–775.

Hackenberg M, Barturen G, Carpena P, Luque-Escamilla PL, Previti C, and Oliver JL (2010). Prediction of CpG-island function: CpG clustering vs sliding-window methods. *BMC Genomics* **11:** 327–341.

Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, and Irizarry RA (2012).

Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* **41:** 200–209.

Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation (2012). Penalized solutions to functional regression problems. *Biometrics* **68(3):**774–783.

Leek JT, Johnson WE, Parker HS, Jaffe AE, and Storey JD (2012). *Bioinformatics*, **28(6):**882–883.

Lister R, Pelizzaola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, and others (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315–322.

Laird P (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11:**191–203.

Maher B. (2008). Personal genomes: the case of the missing heritability. *Nature* **456:**18–21 doi:10.1038/456018a.

Manolio TA, Brooks LD, and Collins FS (2008). A Hapmap harvest of insights into the genetics of common disease. *J Clin Invest* **18(5):** 1590–1605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. (2009). Finding the mising heritability of complex diseases. *Nature* **461(7265):**747–754 doi: 10.1038/nature08494.

Qiu W, Baccarelli A, Carey VJ, Boutaoui N, Bacherman H, Klanderman B, Bernard S, Agusti A, Anderson W, Lomas DA and others. (2012). Variable DNA methylation is associated with chronic obstructive pulmonary disease and lung function. *Am J Respir Crit Care Med* **185(4):**373–381.

Streitberg B and Roehmel J. (1986). Exact distribution for permutation and rank tests: an introduction to some recently published algorithms. *Statist Software Newslett* **12:**10–17.

TCGA Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487(7407):**330–337.

Waller LA and Gotway CA (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons Inc.

2. A principal component approach for the detection of unknown substructure in DNA methylation data

Wai-Ki Yip[1], Martin Aryee[2,3], Hannah R Elliott[4], Nan Laird[1] and Christoph Lange[1]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.
[2]Massachusetts General Hospital, Charlestown, MA 02129, USA.
[3]Harvard Medical School, Boston, MA 02115, USA.
[4]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK.

Abstract

Undetected data substructure can be a major source of confounding in genetic studies, potentially leading to spurious results. The recent deluge of association studies using DNA methylation measurements stresses the need for a methodology that is able to detect substructure in such data and to adjust for it. Here, we propose an approach based on principal coordinates analysis to identify the set of methylation sites that can be used to differentiate between subgroups. Our result shows that only a small subset of all measured methylation sites is needed to identify the substructure. We demonstrate using the Southall And Brent REvisited (SABRE) cohort and a colorectal cancer data set from The Cancer Genome Atlas (TCGA); and perform a simulation study to show that the method can identify the substructure if a moderate shift in the mean occurs (about 0.3 in β-value), even if only a small percentage of the methylation sites are affected.

2.1 Introduction

Genome Wide Association Studies (GWAS) have been very successful in identifying genetic variants of many complex diseases. However, not all heritability has been accounted for (Manolio et al., 2009). Researchers have started looking into DNA methylation as the source of some of the missing heritability. However, undetected substructure of the data could confound observations. A good statistical method based on DNA methylation measurements is therefore required to identify and adjust for data substructure in association tests.

With the current biomedical technologies such as the Infinium HumanMethylation450 BeadChip (Illumina, 2011), we can now interrogate more than 450,000 methylation sites per sample at single-nucleotide resolution enabling researchers to perform epigenome-wide association studies (EWAS). The data at any given methylation site represents the

average methylation level in the sample cell population. The measurement is known as a β-value. This is a continuous value bounded by 0 and 1 which corresponds to the ratio of methylated signal divided by the sum of the methylated and unmethylated signals.

Here, we propose a simple exploratory and visual statistical method to identify substructure and potential methylation sites that define the subgroup differences on data generated using BeadChip arrays. We will refer to the method as Methylation Distance Analysis (MDA). Examples of substructure in population are ethnic subgroups such as Caucasian, Asian or African. Since these subgroups are usually not known or observed, MDA can be used to identify them and allow adjustment be made to the analysis. Thus, MDA reduces spurious conclusions in DNA methylation analyses. In addition, MDA can be used to identify methylation sites that are associated with phenotypes such as subgroups of cancer and normal cells. Additional research can then be done on these sites to determine the biological mechanisms involved.

The MDA method has similarities with the Eigenstrat approach that is commonly used in genetic analysis (Price et al., 2006). The Eigenstrat method applies principal components analysis (PCA) to the sample covariance of the single nucleotide polymorphism (SNP) genetic data. If there is population substructure in the data, the scatterplot of the principal components will separate samples into different clusters. However, the Eigenstrat method is developed for genetic/SNP data. Recent discovery has shown that DNA methylation is also related to population substructure (Fraser, 2012). Thus, we need a similar statistical method to detect and adjust for population substructure in EWAS. The proposed method can be used to compare with a recent research that uses PCA of Genome-Wide DNA methylation data to account for population stratification (Barfield et al., 2014).

The MDA method does not use the sample covariance of the measured methylation β-value. We first define the DNA methylation distance based on the β-values between two samples relative to a selected set of methylation sites. Then, we apply a multi-dimensional scaling (MDS) technique similar to PCA known as principal coordinates analysis (PCoA) (Gower, 2005) to the methylation distance matrix for all samples. Then, each sample is plotted based on the first and second principal components. Similar to the Eigenstrat method, we look for patterns of clustering in the scatterplots of the first two principal components from the PCoA as we vary the selected set of methylation sites.

The number of methylation sites measured on arrays is large and not all of them are necessarily related to the underlying substructure. Thus, including all the methylation sites in the analysis will add too much noise and is consequently not desirable. The proposed method uses a simple variable selection strategy to identify potential informative methylation sites. Only the β-values from this selected set of methylation sites are used in calculating the DNA methylation distance matrix.

To demonstrate the efficacy of the method, we apply it to two different data sets. First, MDA is used to analyze a selected subset of the SABRE cohort data set from the United Kingdom. We want to assess how well MDA identifies the two ethnic subpopulations in the selected SABRE cohort. After applyng MDA, we see the two ethnic groups separate clearly into two clusters in the scatterplots of specific selected sets of methylation sites. Second, MDA is used to analyze a colorectal cancer data set from The Cancer Genome Atlas (TCGA) url: http://cancergenome.nih.gov. In this case, we are exploring the dataset's substructure and find that the cancerous and normal samples separate into clusters in the scatterplots.

MDA is an exploratory analysis tool, ideal for hypothesis generation. The list of informa-

tive methylation sites identified to define the subgroups is not necessarily related to the underlying subgroups. The main result is shown in a scatterplot. Users need to apply their judgement to visually identify the distinctive clusters and associate them with subgroups under consideration. Although these clusters may not be attributed to any known subgroups, user can still adjust for them to avoid confounding. Unfortunately, there is no specific statistic that we can use for testing the existence of clusters or the membership of a sample to a cluster. With MDA, researchers can begin to identify methylation sites that are potentially associated with the underlying substructure and can adjust for them in DNA methylation analysis. This will lead to a reduction in spurious conclusions and a better understanding of the role of DNA methylation in the differentiation of subgroups.

2.2 Approach

The intuitive idea behind the method is that there are biological differences in some of the methylation sites of individuals belonging to different subgroups. This biological difference is reflected in the DNA methylation measurement which is measured as a β-value. Individuals in the same subgroup are closer to one another in terms of DNA methylation distance and people in a different subgroup should be farther away.

The first step is to define the DNA methylation distance, md_{ij}, between individual i and individual j:

$$md_{ij} = \sqrt{\sum_{k} (\beta_{ik} - \beta_{jk})^2} \tag{2.1}$$

where k is the k^{th} CpG sites in the p-elements subset (K) of all measured methylation sites under consideration. Similar to other distance measures, this is the Euclidean distance in a p-dimensional space with p is the number of methylation sites selected by a

variable selection algorithm. The algorithm separates the samples of each methylation site into two subgroups. It does not matter if the underlying structure has more than two subgroups. This separation tries to capture information about the biggest difference among subgroups. The methylation site is considered as informative if the mean of the methylation β-values of the two subgroups is greater than or equal to a predetermined threshold, δ. The proposed algorithm iterates with δ starting from 0 until the informative set is empty. For our SABRE cohort data, when δ is set as 0, all 450,725 measured methylation sites will be considered as informative. Thus, using this informative candidate set, we can calculate the symmetric methylation distance matrix (MDM), for each δ, for all n individuals.

$$\begin{pmatrix} 0 & md_{12} & \dots & md_{1n} \\ md_{21} & 0 & \dots & md_{2n} \\ \dots & & & \\ md_{n1} & md_{n2} & \dots & 0 \end{pmatrix} \tag{2.2}$$

Note that $md_{ii} = 0$ for $i = 1, \dots, n$ and $md_{ij} = md_{ji}$ for all i and j. We define the set $K_\delta = \{$ methylation sites $||$difference of mean β-values of the two subgroups $| \geq \delta$ for some predetermined $\delta\}$ and $p = \|K_\delta\|$, the number of elements in the set K_δ. When δ is 0, all measured methylation sites are included.

2.3 Methods

In most experiments, samples are collected from the study population. Individuals may belong to certain unknown subgroups which could affect the analysis results. We do not know how many subgroups there are or to which subgroups each sample belongs. We apply a variable selection algorithm to provide us with a set of informative methylation sites. Then, we perform the principal coordinate analysis based on this informative set of sites as follows:

2.3.1 The MDA Procedure

1. Model all methylation sites with a mixture of two Gaussian models (using the R package Mclust (Fraley, 1998)).
2. Find the absolute differences of the mean of the two Gaussian distributions from the mixture for each methylation site.
3. Set $\delta = 0$.
4. Define the set K_δ by selecting all methylation sites whose difference in means computed in step 2 is greater than or equal to δ.
5. Create MDM based on K_δ.
6. Perform a Principal Coordinate Analysis (PCoA) on the MDM created using the R package labdsv (Roberts, 2013) url: http://CRAN.R-project.org/package=labdsv.
7. Check visually the scatterplot of samples using the first two principal components to see if identifiable clusters appear.
8. Change δ by a small increment, e.g. 0.05.
9. Repeat step 4-8 until the set K_δ is empty.
10. Identify the clearest plot with distinctive clusters.

At this stage the set K_δ corresponds to the plot identified in step 10 contains all the methylation sites needed for adjustment. Note that this is a visual procedure and one has to apply his/her judgment for proper identification. You may not detect any clusters at all. It all depends on how close one subgroup is to the other in terms of DNA methylation distance.

2.3.2 Motivating Data Set 1: The SABRE Cohort

The SABRE cohort is a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origin. A detailed description of the cohort profile can be found in (Tillin et al., 2012).

We took a subset of 192 male samples from the SABRE cohort that were previously used in a study to assess differences in smoking associated DNA methylation patterns in South Asians and Europeans. A detailed description of the study and sample selection can be found in (Elliott et al., 2014). The reason for choosing this data set is that it has two well defined major ethnic groups while many of the individual baseline characteristics such as age are similar.

Data are collected using Illumina HumanMethylation450 BeadChips. Methylation levels for the sites are measured in β-values and are preprocessed and normalized using quantile normalization. The ComBat function in the R package sva (Leek et al., 2012) is used to correct for batch effects. After data cleaning, a total of 450,725 methylation sites in 189 samples are used in this analysis. The purpose of this investigation is to find methylation sites that identify the known population substructure in the data.

2.3.3 Motivating Data Set 2: The TCGA Cancer Data

The colorectal cancer data set is a publicly available from the Cancer Genome Atlas (TCGA) website. The data are collected using Illumina Human Methylation450 BeadChips. There are a total of 329 samples from cancer patients. We used only 76 matched samples from the same patients - one sample from solid tumor and a corresponding sample from normal tissue. These matched samples are processed in 3 separate batches. We

used the ComBat function in the R package sva (Leek et al., 2012) to correct for batch effects. However, normalization is not done with the samples. After data cleaning, a total of 385,885 methylation sites are used in the analysis. The purpose of this investigation is to explore any underlying substructure in the data.

2.3.4 Simulation Study

In order to understand the capability of MDA to detect the shift of mean β-values, we performed an additional simulation study. We took the European subgroup of the SABRE cohort which has a total of 95 samples as our base group. We created another subgroup of 95 samples by first bootstrapping the sample values for each methylation site from the base group and then added a constant β-value of a randomly selected number of methylation sites to simulate the effect. We chose the bootstrapping methodology because we did not want to make assumption about the distribution of these β-values. We picked 1000, 100, 10 and 2 associated methylation sites randomly from all measured sites for the simulation study. For each scenario, we added or subtracted to each of these associated sites by a specific β-values amount of 0.5, 0.3 and 0.1. To maximize the efffect, we checked the mean β-values of each bootstrapped associated sites. If the mean β-value of that site was over 0.5. we subtracted and if the value was less than 0.5, we added the predetermined constant amount. The final β-value was truncated at 0 and 1.

2.4 Result

2.4.1 MDA method applied to the SABRE cohort

We first applied the R program Mclust to model a mixture of two Gaussian models for all methylation sites. Then, we computed the differences for the mean of these two Gaussian models for all methylation sites. The largest absolute difference was greater than or equal to 0.9.

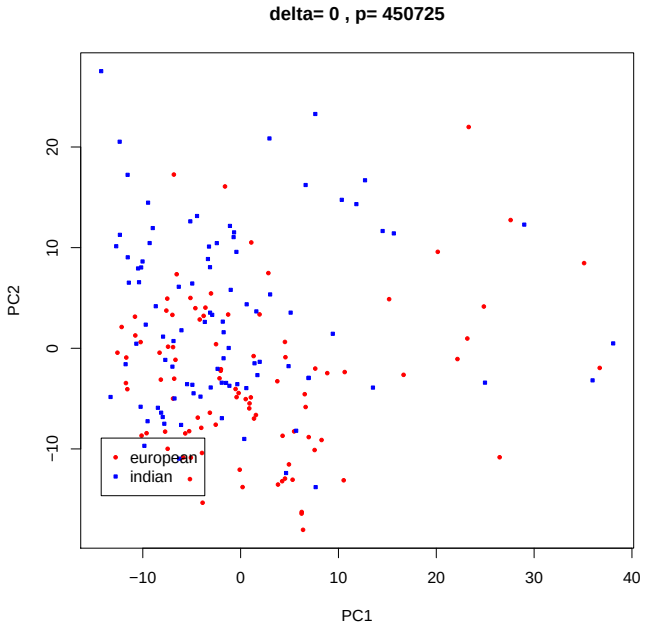

Figure 2.1: SABRE: Using all 450725 methylation sites, i.e. with $\delta \geq 0$, the two subpopulations are mixed together.

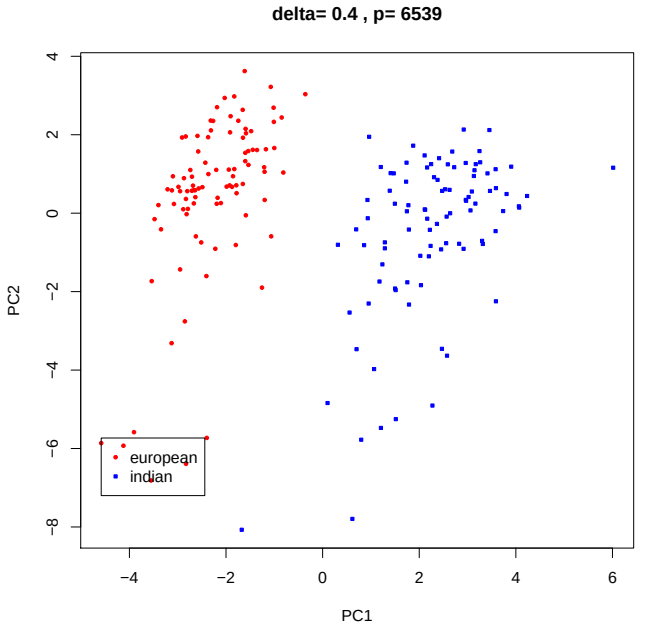

Figure 2.2: SABRE: With $\delta \geq 0.4$, 6539 methylation sites are identified. The two clusters representing the two ethnic groups are now discernible.

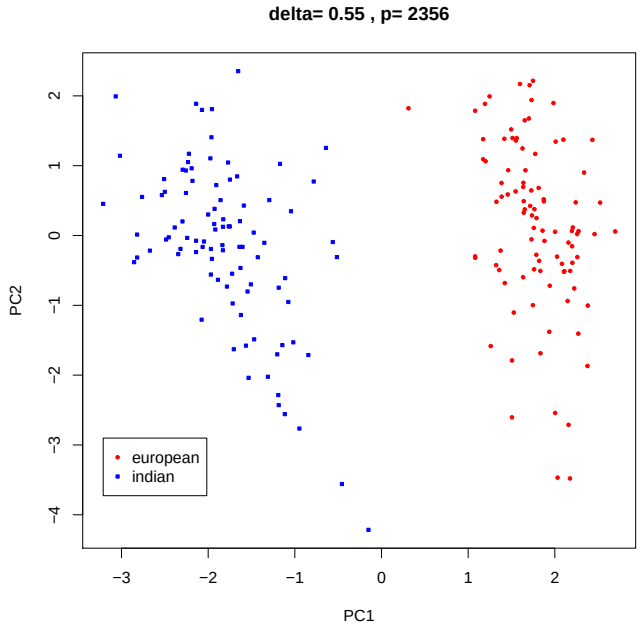

Figure 2.3: SABRE: With $\delta \geq 0.55$, 2356 methylation sites are identified. This seems to be the clearest separation.

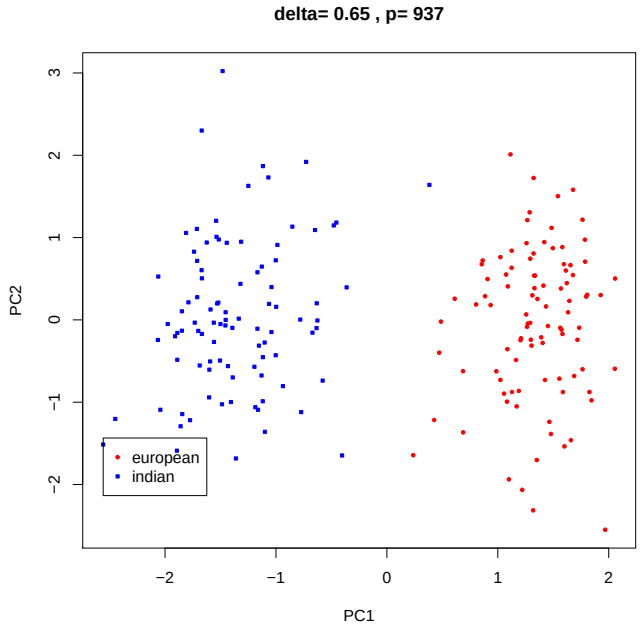

Figure 2.4: SABRE: With $\delta \geq 0.65$, 937 methylation sites are identified. The two clusters start to move closer together.

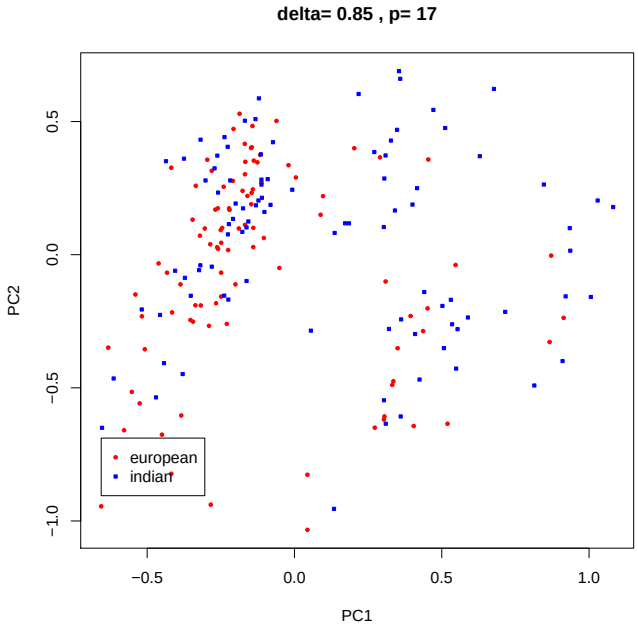

Figure 2.5: SABRE: With $\delta \geq 0.85$, only 17 methylation sites are identified. The two clusters are now mixed together again.

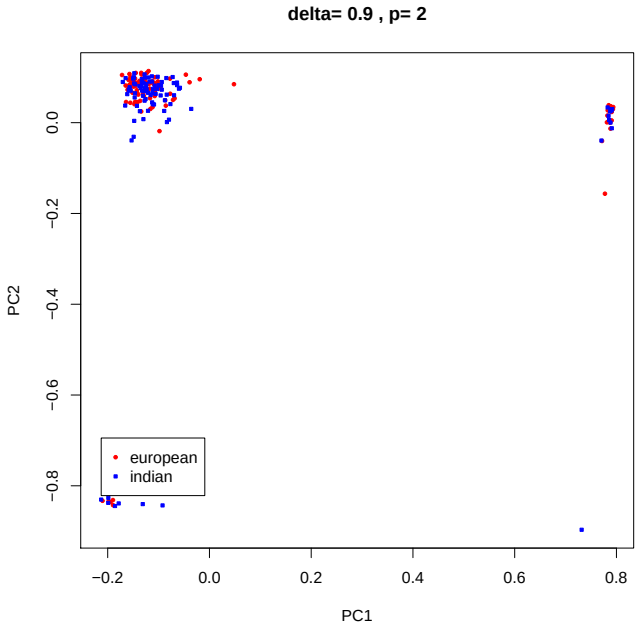

Figure 2.6: SABRE: With $\delta \geq 0.9$, 2 methylation sites are identified. There seems to be four clusters here. However, this pattern is not consistent.

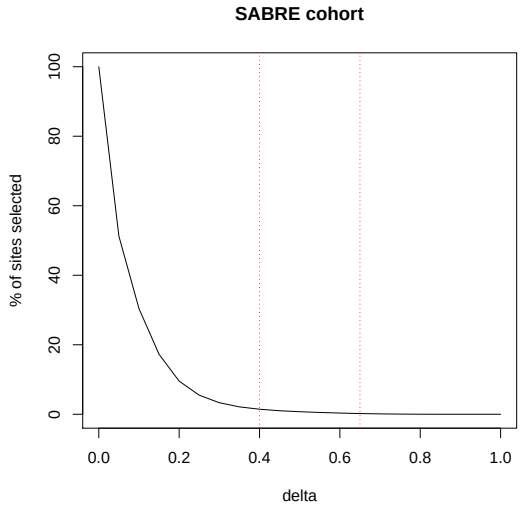

Figure 2.7: SABRE: The % of methylation sites selected for the set K_δ is plotted against the specific delta δ. The clustering patterns are observed between the two "red" dotted lines.

We created the sets K_δ by varying δ starting from 0 to 0.9 with an increment of 0.05. For each set K_δ, we computed the distance matrix MDM and analyzed it using PCoA. The sequence of scatterplots from Figure 2.1 to Figure 2.6 showed the clustering patterns by varying the value of δ. For brevity reason, only the scatterplots with interesting patterns were displayed. The entire sequence of scatterplots could be found in Appendix A.2.1.

Summarizing the results, no specific patterns or clusters appeared until $\delta \geq 0.4$ with 6539 methylation sites used in PCoA (Figure 2.2). The clearest plot was displayed when $\delta \geq 0.55$ with 2356 methylation sites (figure 2.3). However, the two clusters started to merge into one when $\delta \geq 0.65$ with 937 methylation sites used (Figure 2.4). They were completely merged with $\delta \geq 0.85$ with 17 methylation sites used (Figure 2.5). A visual summary of the result was given in Figure 2.7.

2.4.2 MDA method applied to the TCGA dataset

As with SABRE data, we first appplied MClust to obtain the 2 Gaussian models for each methylation site. The maximum absolute difference was 0.82. Again, only plots with

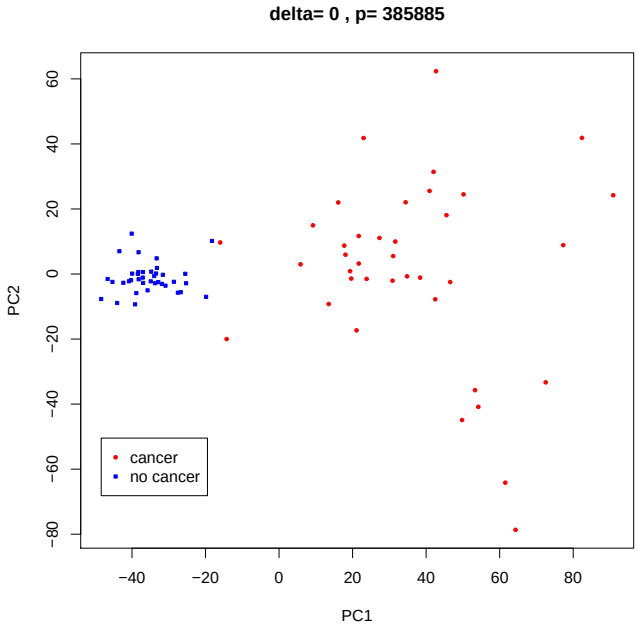

Figure 2.8: TCGA: Using all 385,885 methylation sites i.e. with $\delta \geq 0$, it shows two distinctive subgroups: a dense non-cancerous cluster on the left and a more scattered cancerous cluster on the right. This pattern persists for almost all δ.

interesting patterns were shown here and the entire sequence of scatterplots could be found in Appendix A.2.3.

Summarizing the results, the specific pattern persists for δ ranging from 0 to 0.65. We surmise that the cancer signal might come from a small number of methylation sites and they were so strong that they overcame all other noises. A visual summary of the result is given in Figure 2.11.

2.4.3 Simulation study

The purpose of the simulation study was to explore the range of mean β-value shifts in the methylation sites to see if MDA could identify them. With the new sample, we applied MDA and visually examined the plots.

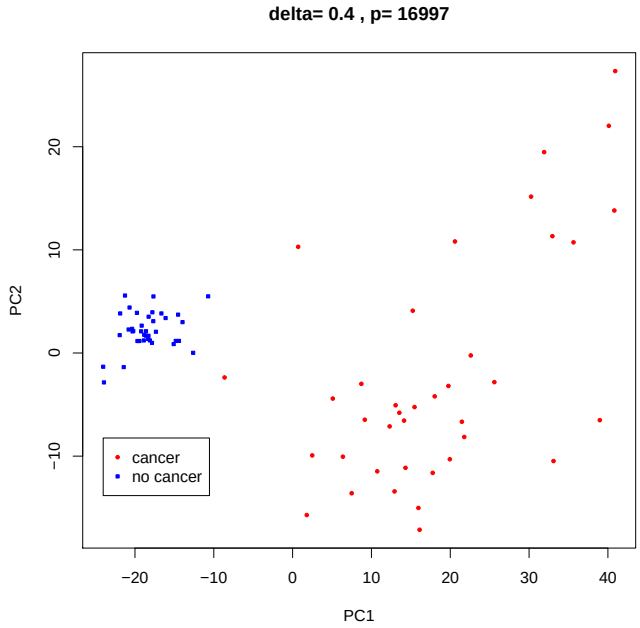

Figure 2.9: TCGA: This seems to have the clearest separation using 16,997 methylation sites with $\delta \geq 0.4$.

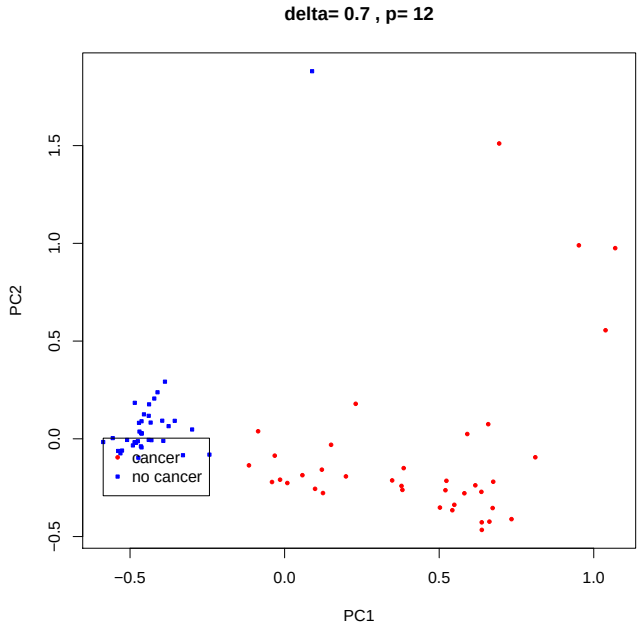

Figure 2.10: TCGA: Using only 12 methylation sites with $\delta \geq 0.7$, the clusters are no longer distinguishable unless the labels are known.

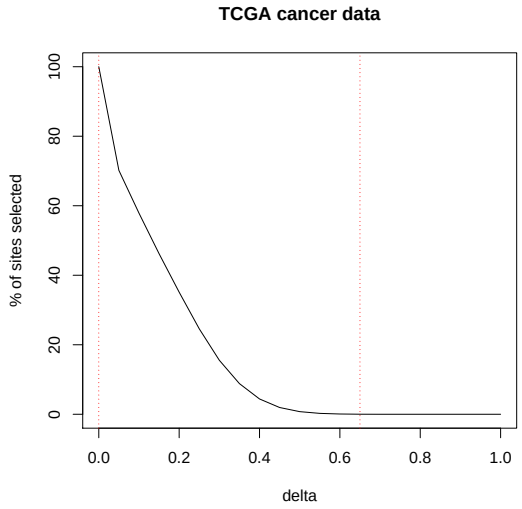

Figure 2.11: TCGA: The % of methylation sites selected for the set K_δ is plotted against the specific delta δ for the TCGA colorectal cancer data. The clustering patterns are observed between the two ″red″ dotted lines.

Table 2.1: Result of the simulation study.

no of sites	Shift in methylation β-values		
	0.5	0.3	0.1
1000	Samples separate into two clusters when $0.38 \geq \delta \geq 0$.	Samples separate into two clusters when $0.3 \geq \delta \geq 0.26$. It merged into one when $\delta \leq 0.22$.	No pattern obs.
100	Samples separate into two clusters when $0.38 \geq \delta \geq 0.5$. It merged into one when $\delta \leq 0.36$.	No pattern obs.	No pattern obs.
10	No pattern obs.	No pattern obs.	No pattern obs.
2	No pattern obs.	No pattern obs.	No pattern obs.

As shown in Table 2.1, MDA identified the substructure only when it captured a substantial percentage of the associated sites and when the differences in β-values were large (> 0.3). This was completely consistent with what we noticed in the SABRE cohort and the TCGA data analysis. If the signal was weak, we would not be able to discern any pattern at all. If the signal was moderate (≤ 0.3 and > 0.15), we could observe the clustering pattern within a narrow range of threshold values. If the signal was strong, the clustering pattern could persist even with all measured methylation sites included.

2.5 Discussion and Conclusion

MDA has been presented as a method to identify population substructure in methylation data. We demonstrated MDA method's efficacy by applying it to our sample data, the SABRE cohort. With only 2356 methylation sites (0.5% of measured sites), MDA separated the two ethnic groups completely. It would be worthwhile to investigate the biological role of these methylation sites in the differentiation of ethnic groups. Furthermore, MDA was tolerant of a lot of noise. The separation pattern in SABRE continued to hold until after 6534 methylation sites were included. Most likely, many of these methylation sites were not related to population substructure. One caveat was that when the number of methylation sites used were very small, some spurious clustering might occur. One should always examine all scatterplots for all δs before making any judgement on the patterns of clustering.

We obtained similar results when MDA was applied to the TCGA data set. In this case, we only needed 59 methylation sites to establish the pattern where the cancerous and non-cancerous samples separated into two clusters. This pattern of clustering persists even when we included the entire 385,885 measured sites, possibly because a large fraction of these sites were hypomethylated as in most cancer samples.

In the simulation study, we showed that MDA was capable of detecting the associated sites if their effect measured in shift in β-values were strong and plentiful enough (greater than 0.3 and in more than 100 associated sites). If there were few associated sites, MDA could not identify the substructure introduced in the samples. This finding was consistent with our results in the SABRE cohort and TCGA data analyses.

Currently, there is no standard method on how to adjust for unobserved confounding in DNA methylation data. Recently, a Principal Component based method to account for population stratification in DNA methylation studies (Barfield et al., 2014) is proposed. However, it requires users to do pruning. The pruning can be done by looking at correlation between methylation sites and relevant genetic variants. In addition, certain location information could be used to identify these informative sites. MDA identifies these informative sites differently. Instead of correlation, MDA uses a distance function to capture the differences. The sites identified can be used in a PCA analysis as detailed in Barfield's paper. For future research, we can compare the power of MDA and other pruning methods in the adjustment of population substructure. Surrogate Variable Analysis (SVA) (Leek et al., 2012) is another method proposed to adjust for unbserved confounding. However, one has to build a model first and sva tries to locate the confounding factors after removing the main effect. MDA does not depend on a model and it is more flexible as the distance function can be modified to search for other differences in DNA methylation profiles besides a mean difference. However, since it does not fit into a model framework directly, it can only be used for exploratory data analysis.

There are numerous benefits in using MDA. Since the outcome is visual, it is easy to show and explain the results once the clustering pattern is identified. If the clustering pattern can be labeled with certain substructure (such as population substructure), one

can adjust for the substructure to avoid spurious association. Although MDA does not provide a statistic and allow one to make inferential statement about an association, it narrows the number of methylation sites to be considered substantially from $\tilde{4}$50,000 to just a few thousand or less. Thus, it allows one to explore the data set and reduces the complexity of finding the relevant informative methylation sites assocaited with certain substructure to a manageable subset of candidates. Then, one can generate hypotheses and design appropriate experiments to establish the association with the candidate set.

The study reported here has limitations. So far, we have only applied our method to two studies which contain distinct and homogeneous subgroups. In practice, the admixture may be a lot more diverse. The clustering pattern may not be clear cut. We don't yet have information about MDA's performance in a more complex situation.

The MDA algorithm has certain drawbacks. First and foremost, it is a visually based method. One has to apply his/her subjective judgment to look for some sort of clustering patterns since there is no statistical tests for any guidance in selecting the optimal number of methylation sites to use. We have observed that if we use a very small number of methylation sites, the cluster patterns are not consistently. If a consistent pattern appears, one has to guess what the substructure that it may attribute to. There is no guarantee that the observed clustering pattern is related to the substructure of interest. MDA may capture some/all variation related to the measure of interested or it may be capturing other sources of variation that you might hope to remove. Thus, MDA is best for exploration and hypothesis generation only at this time. The DNA methylation distance used is best used to detect a shift in mean β-values only and is unlikely to detect other kind of differences. It is also plausible subgroups are associated with changes in variance between groups.

Some of these drawbacks should be investigated in future research. One can explore using other distance functions (such as Manhattan, Canberra) to capture other possible methylation differences besides a simple shift in mean β-values. Instead of using Gaussian mixture models, we can try other data mining algorithms that can give us a better candidate set. The Eigenstrat method uses SNP data to separate ethnic groups. We would like to compare our methods using DNA methylation data and the Eigenstrat method. It will give us information about the relationship between genetic and epigenetic data in the divergence of ethnic groups in human population as explored in (Liu et al., 2010). The SABRE cohort and TCGA data set have been cleaned thoroughly - removal of batch effect through ComBat for both and quantile normalization with the SABRE cohort. These data cleaning techniques could have reduced global effects. We need to investigate the effect of cleaning and normalization method's effect on MDA. Finally, we want to try it on additional methylation data sets and to find out how MDA generalizes to a more complex situation such as applying it to identify subtypes of cancer or cell type compositions in DNA methylation data.

There are additional potential for MDA for used as an association discovery tool. In some situations where the labels of the subgroups are known, MDA can be extended to identify the methylation sites that are associated with each subgroup. However, more research is needed to confirm the validity of this approach.

The MDA method is a simple method to use. Although it is only an exploratory and visual method, it can be used to identify methylation sites that are associated with substructure in the data. Then, adjustment can be made if these methylation sites are used in the association study. Furthermore, this is an additional tool for researchers to explore the role of different genetic mechanisms such as genetic variants and DNA methylation on how human evolves.

Acknowledgements

We are grateful to the SABRE study participants and staff. SABRE is led by Nish Chaturvedi, Alun D Hughes and Therese Tillin. The study was funded at baseline by the Medical Research Council, Diabetes UK, and the British Heart Foundation and at follow-up by the Wellcome Trust [WT082464] and British Heart Foundation [SP/07/001/23603]. Methylation analysis in the SABRE cohort was supported by a Wellcome Trust Enhancement grant [082464/Z/07/C].

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government [NRF-2014S1A2A2028559].

Hannah Elliott is supported by an Oak Foundation post-doctoral research fellowship award. She works in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol which is supported by the Medical Research Council and the University of Bristol [MC_UU_12013/2]. Wai-Ki Yip is supported by the Clinical Epidemiology of Lung Diseases Training Grant [T32 HL 007427].

References

Tillin,T., Forouhi N.G., McKeigue P.M., Chaturvedi N. (2012). Southall And Brent Revisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of Euuropean, Indian Asian and African Caribbean origins, *International Journal of Epidemiology*, **41**, 33-42.

Elliott H.R., Tillin T., McArdle W.L., Ho K., Duggirala A., Frayling T.M., Smith G.D., Hughes A.D., Chaturvedi N., Relton C.L. (2014). Differences in smoking associated DNA methylation patterns in South Asians and European, *Clinical Epigenetics*, **6:4**, 1-10.

Liu J., Hutchinson K., Perrone-Bizzozero N., Morgan M., Sui J., Calhoun V. (2010) Identification of Genetic and Epigenetic Marks Involved in Population Structure. *PLOS One*, **5:10**, e13209.

Fraser H.B., Lam L.L., Neumann S.M., Kobor M.S. (2012) Population-specificity of human DNA methylation. *Genome Biology*. **13:R8**.

Manolio T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461(7265)** 747-53 doi:10.1038/nature08494.

Leek J.T., Johnson W.E., Parker H.S., Jaffe A.E., Storey J. 2012 The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28** 882-882.

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., Reich D. (2006)

Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, **38** 904-909.

Fraley C., Raftery A.E. (1998) MCLUST: Software for Model-based Cluster Analysis, *Technical Report No 342*, Department of Statistics, University of Washington, Seattle, WA, USA.

Gower J.C. (2005) Principal Coordinates Analysis. *Encyclopedia of Biostatistics*. Vol. 6, ed. Armitage P., Colton T. 2nd edn. West Sussex, England, 4223-4237.

Roberts D.W. (2013) labdsv: Ordination and Multivariate Analysis for Ecology, *R package*, version 1.6-1.

Illumina (2011) Illumina HumanMethylation450 BeachChip Product Information.

Barfield R.T., *et al*, (2014) Accounting for Population Stratification in DNA Methylation Studies *Genetic Epidemiology*, **38** 231-241.

3. STEPP Subpopulation Analysis for Continuous, Binary and Count Outcomes

Wai-Ki Yip[1], Marco Bonetti[2], Benard F. Cole[3], William Barcella[4], Xin Victoria Wang[1,5], Ann A. Lazar[6], and Richard D. Gelber[1,5]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA
[2]Department of Policy Analysis and Public Management, Bocconi University, Milan, Italy
[3]Department of Mathematics and Statistics, University of Vermont, Burlington, VT, USA
[4]Department of Statistical Science, University College London, London, UK
[5]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[6]Department of Preventive and Restorative Dental Sciences, University of California, San Francisco, CA, USA

Abstract

Medical research is increasingly focused on personalizing the care of patients with chronic diseases, especially given recent advancements in the availability of biomarkers of treatment sensitivity. Personalized medicine requires investigating how patient characteristics, including novel biomarkers, modify the effect of current treatment modalities. A better understanding of the interaction between treatment and patient specific prognostic factors will enable practitioners to expand the availability of tailored therapies with the ultimate goal of improving patient outcomes. The Subpopulation Treatment Effect Pattern Plot (STEPP) approach was developed to allow researchers to investigate the heterogeneity of treatment effects on survival outcomes across increasing values of a (continuously measured) covariate, such as biomarker measurement.

We extend the STEPP approach to continuous, binary and count outcomes which can be easily modeled using generalized linear models (GLM). With the extension in STEPP, these new kinds of treatment effects within subpopulations defined with respect to a covariate of interest can be estimated, and the statistical significance of any observed heterogeneity of treatment effect is assessed using permutation tests. The desirable feature that commonly used models are applied to well-defined patient subgroups to estimate the treatment effects is retained in this extension. We conduct a simulation study to confirm that the proper type I error rate is maintained when the outcome is independent of the covariate of interest. The statistical methods are applied to motivating data from the Asprin/Folate Polyp Prevention Study, a clinical trial evaluating the effect of oral aspirin, folic acid, or both as a chemoprevention agent against colorectal adenomas (including lesion). The R software package (**stepp**) has been extended to handle continuous, binary and count data using Gaussian, binomial and Poisson models.

3.1 Introduction

Results from randomized clinical trials (RCT) provide the foundation of evidence-based medicine. RCTs often compare the benefits of two competing therapies, and they may provide evidence to establish optimal treatment combinations. The measurement of effectiveness is typically based on the entire cohort of patients enrolled in the study. However, the magnitude of the treatment effect may be heterogeneous among patient subpopulations (e.g., across different age groups). Instead of the traditional one-size-fits-all treatment recommendation, understanding the interaction between treatment and covariates may provide the information necessary to allow physicians to customize treatment to individuals, thus maximizing the treatment benefits.

A common approach in that direction is the examination of treatment effect within subsets of the patient population. Performing subgroup analysis is in general a challenging task (Lagakos, 2006; Wang, Lagakos and Ware, 2007). Traditionally, patients are divided into subgroups according to median, quartiles or other convenient cut-points of one or more covariates of interest, and treatment comparisons are then performed within each subgroup. Unfortunately such cut-points, while convenient, do not necessarily identify clinically important subgroups; furthermore, they might fail to detect complex associations such as non-linear or bimodal interactions. Treatment-covariate interactions for survival data can be analyzed using regression methods such as the Cox proportional hazards model (Cox, 1972) or the cumulative incidence models of Fine and Gray (Fine and Gray, 1999; Gray, 1998). However, such models require one to define a functional form for the treatment-covariate interaction.

The subpopulation treatment effect pattern plot (STEPP) method was developed as an alternative approach to identify treatment-covariate interactions (Bonetti and Gelber, 2000, 2004; Bonetti, Zahrieh, Cole and Gelber, 2009). STEPP is primarily a graphical tool designed to help researchers explore the potential heterogeneity of treatment effect and facilitate the interpretation of estimates of treatment effect derived from different but possi-

bly overlapping subsets of patients defined by the values of a continuous covariate such as a risk index. First, STEPP divides the population into overlapping subpopulations; second, STEPP estimates the treatment effects for each treatment in each subpopulations. Finally, these effects are plotted against the covariate of interest. The method is aimed at determining whether the magnitude of the treatment effect changes for different values of the covariate used to define the subpopulations. STEPP has the advantage of making no a priori assumption regarding the pattern of interaction and thus has the potential to elucidate complex associations. Furthermore, by allowing subpoulations to overlap, the estimated treatment effect in a particular covariate interval utilizes information from adjacent observations. Importantly, STEPP uses well-known methods to estimate treatment effects within well-defined groups of patients.

STEPP was developed for the analysis of time-to-event data (Bonetti and Gelber, 2000, 2004; Bonetti, Zahrieh, Cole and Gelber, 2009; Lazar, Cole, Bonetti and Gelber, 2010). An R (RCORE, 2008) software package (**stepp**) is available through CRAN `<cran.us.r-project.org>` (Yip, 2011). Users can analyze the following commonly used measures of treatment effects: difference in Kaplan-Meier estimates of survival functions at specific time points, difference in cumulative incidence of a disease specific event in the presence of competing risks; and hazard ratio estimates based on observed minus expected estimation, all with a single end point.

The method has been applied successfully to analyze a number of clinical trials using the R package. Examples include the Breast International Group (BIG) 1-98 randomized clinical trial evaluating adjuvant therapy with letrozole versus tamoxifen for postmenopausal women with hormone-receptor-positive breast cancer (Viale et al., 2008), and the evaluation of the effect of different durations of adjuvant chemotherapy (3 vs 6 courses of CMF) for young women with breast cancer (Colleoni et al., 2002). In both of these studies, the STEPP analysis revealed that the magnitude of treatment effect varied with the covariate of interest.

More generally, however, researchers often perform clinical trials with non-survival endpoints. For example, researchers may be interested in evaluating the risk of certain diseases based on exposure to factors such as harmful chemicals. The aim here is to extend the STEPP approach and software to analyze non-survival data. We restrict our attention to the analysis of continuous, binary and count data which can be modeled by GLM models - the Gaussian model with the identity link, the binomial model with the logit link, and the Poisson model with the log link respectively. Using a modeling approach, investigators have the flexibility of incorporate other covariates for predicting the outcome.

As an illustration, we applied GLM STEPP to motivating data from the Aspirin/Folate Polyp Prevention Study, which we briefly describe in section 3.2. (Baron et al., 2003). Our analysis explored the potential interaction between aspirin treatment and age on the occurrence of colorectal adenomas (including lesion)(AD).

In section 3.3, we describe our proposed extension. In particular, for statistical inference we assess the interaction effects by computing permutation p-values based on several statistics. In section 3.4, we briefly summarize the main results from a simulation study aimed at confirming the maintenance of proper type I error rate with the statistics, and show the result of the application of the methods to the Aspirin/Folate Polyp Prevention study. We close with some discussion in section 3.5.

3.2 Motivating Data: The Aspirin/Folate Polyp Prevention Study

The Aspirin/Folate Polyp Prevention Study was a randomized, double-blind, placebo-controlled trial of the efficacy of oral aspirin, folic acid, or both to prevent colorectal AD. Our analyses here are confined to the aspirin component of the study. There were 1,121 participants randomized to three aspirin groups (placebo, 81 mg/day and 325 mg/day). Participants were followed for three years, and then underwent colonoscopy. The pri-

mary endpoint was the occurrence of any pathologically confirmed AD. A total of 1,084 participants underwent colonoscopic follow-up at three years. The original findings of the aspirin analysis (Baron et al., 2003) concluding that low-dose aspirin had a moderate chemopreventive effect on AD in the large bowel. We used STEPP to investigate whether the magnitude of the treatment effect is similar across subpopulations defined by patient age. In particular, we are interested in exploring the interaction effects in the three pairwise comparisons of treatment arms, namely placebo vs 81 mg of aspirin, placebo vs 325 mg of aspirin, and 81 mg vs 325 mg of aspirin.

The first analysis is presented in section 3.4.2 , whilst the second and third analyses are included in the appendices A.3.1 and A.3.2.

3.3 Methods

STEPP is a graphical method developed to address some of the concerns associated with traditional subgroup analysis. The main advantages of STEPP are the fact that it requires few assumptions and that it provides a graphical display to show potentially complex interactions, thus assisting clinicians in the interpretation of results. Also, treatment effect estimates are based on widely used methods applied to well defined groups of patients.

While traditional statistical methods for subgroup analysis divide the population into disjoint subgroups, STEPP takes a different approach in that it constructs overlapping subpoulations along the continuum of the covariate of interest, thereby improving the precision of the estimated treatment effect within the subgroups in a smoothing-by-binning manner (Bonetti and Gelber, 2000).

STEPP can be implemented via two specific patterns of subpopulations: the "sliding window" pattern and the "tail-oriented" pattern. In a clinical trial, we consider n patients being assigned randomly to one of the two treatments. A subpopulation of the sliding window pattern is defined by two cutoff values $[z_{min}, z_{max}] \subset \Re^2$ so that patients who are

randomly assigned to one of the two treatment arms with the covariate value (Z) between these two cutoffs are chosen as part of this subpopulation. The subpopulations of the tail-oriented pattern form an increasing nested sequence of subpopulations, with the last one containing all patients, also using well-defined cut points (Bonetti and Gelber, 2000).

Since most applications of STEPP use the sliding window pattern, in the remainder of this paper we focus on that construction. The sliding window pattern is implemented by using two window parameters chosen by the investigator - the number of patients per subpopulation ($r2$) and the largest number of patients in common between two consecutive subpopulations ($r1$) to construct subpoulations based on the value of a covariate of interest (or a risk score). One then slides the window forward by replacing $(r2 - r1)$ individuals with new individuals with higher value of the covariate Z (assuming there are no ties). These two numbers determine the number of subpopulations. Asymptotically, as $n \longrightarrow \infty$, the number of subpopulations approaches $\lfloor 1 + \frac{(n-r2)}{r2-r1} \rfloor$, where the last subpopulation contains a proportion of patients at most equal to $r2/n$.

3.3.1 Treatment Effects for Continuous, Binary and Count Data

For survival analysis, the treatment effects may be defined as differences in survival at a fixed time point between the two treatment arms (e.g. hazard ratios, or cumulative incidence). Here, we consider the STEPP when using continuous, binary and count data which can be modeled using GLM models (Gaussian, binomial and Poisson). Table 3.1 contains the detailed definitions of treatment effect in absolute and relative scales. A GLM model is fitted for each subpopulation which is used to estimate the treatment effect for that subpopulation. Investigators should examine interaction effects in both scales as they may be statistical significant in one scale but not the other. Also, note that the overall treatment effect is in general not a linear function of the subpopulation specific treatment effects.

Within the K subpopulations $P_j, j = 1, \cdots, K$, constructed using the sliding window pat-

tern, a vector $(\hat{\theta}_1, \cdots, \hat{\theta}_K)$ of estimates for the treatment effects based on the fitted GLM model are produced. Similar to the STEPP analysis for survival data, the final result of the STEPP procedure are plots of the treatment effect estimates across the subpopulations, against the median values of the covariate of interest in those subpopulations. The estimated treatment effects provide a graphical presentation of the heterogeneity of treatment effect according to the value of the baseline covariate (Z). For GLMs, the outcome measures of interest for each treatment arm within each subpopulation are first plotted on the vertical axis against the subpopulation specific median values of the covariate that define the subpopulations. An example of such plot can be found in Figure 3.1.

A second plot shows the differences of the treatment outcome measures within each subpopulation by plotting these differences against the same median values. Finally, a third plot shows the ratios of the treatment outcome measures within each subpopulation. The corresponding simultaneous confidence intervals are also provided in the second and third STEPP plots. Examples of these two additional plots can be found in Figure 3.2 and Figure 3.3. Note that the points corresponding to the different treatment effect estimates are only joined for ease of visualization.

If one would like to use GLM STEPP to examine the treatment effects while adjusting for covariates, certain restrictions would apply. Adjusting for covariates imposes the scale on the treatment effects that one can use and so only certain STEPP plots display the adjusted effects properly. For the Guassian model, it would be the second STEPP plot showing the effect differences; for binomial model, it would be the third STEPP plot showing the odds ratio; and for Poisson model, it would be the third STEPP plot showing the relative risks. One should interpret other STEPP plots in this context cautiously.

3.3.2 Inference

In order to properly interpret the three STEPP plots, we associate with each of the treatment effect plots a p-value. For each subpopulation P_j, an estimate $\hat{\theta}_j$ of treatment effect

is computed. Such treatment effect estimates are clearly correlated, as there are a number of patients in common between neighboring subpopulations. For testing the absence of interaction, the following null hypothesis is of interest;

$$H_0 \colon \theta_1 = \theta_2 = \cdots = \theta_K \tag{3.1}$$

We use a permutation approach (see, e.g. Peasarin, 2001) to estimate the covariance matrix of the $\hat{\theta}_j$s and to produce p-values for several possible test statistics. Simultaneous confidence intervals around the collection of estimators are also constructed.

For example, we use the following logistic model to estimate the risk: $logit(p) = trt \times \alpha + X\beta$ where trt is the treatment indicator and X are other covariates. By fitting the binomial GLM to both treatment groups for each subpopulation, we can compute the difference in risk using the formula $\hat{\theta}_j = \hat{p}_{A,j}(t) - \hat{p}_{B,j}(t)$ where $\hat{p}_{G,j}(t)$ is the GLM estimate of the risk within treatment group G inside subpopulation P_j. The $K-$ dimensional vector of estimates, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_K)$ and their approximate variance-covariance matrix $\hat{\Sigma}$ can be estimated consistently from the data via permutation.

The following test statistics can be considered to evaluate the treatment effect in the absolute scale. They are designed to detect different kinds of departures from the null.

1. A supremum statistic on the absolute scale:

$$T_1 = max_{j=1,\cdots,K} |\hat{\theta}_j - \hat{\theta}_{ALL}| / \hat{\sigma}_j, \tag{3.2}$$

where $\hat{\sigma}_j = \sqrt{\hat{var}(\hat{\theta}_j - \hat{\theta}_{ALL})}$ with $\hat{\theta}_{ALL}$ the treatment effect for all subjects taken together.

The supremum statistic is meant to detect a significant sharp departure of the difference from the overall effect.

2. A quadratic form statistic on the absolute scale:

$$T_2 = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ALL})^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ALL}) \tag{3.3}$$

where $\hat{\boldsymbol{\theta}}$ is the $K-$dimensional vector of treatment effect estimates for the K subpopulations; $\hat{\boldsymbol{\theta}}_{ALL}$ is a $K-$dimensional vector with all elements equal to the estimated treatment effect $\hat{\theta}_{ALL}$; amd $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix of the vector $(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{ALL})$.

The quadratic form statistic is meant to detect global departures from the overall effect.

3. A homogeneous association statistic on the absolute scale:

$$T_3 = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_W)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_W) \tag{3.4}$$

where $\hat{\boldsymbol{\theta}}$ is the $K-$dimensional vector of treatment effect estimates for the K subpopulations; $\hat{\boldsymbol{\theta}}_W$ is a $K-$dimensional vector with all elements equal to a weighted average of the estimated treatment effects $\hat{\theta}_j$. Such weighted average can be taken in many different ways, and we pick weights equall to $1/\hat{\sigma}_j$, where $\hat{\sigma}_j$ is the estimated standard deviation of $\hat{\theta}_j$. $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix of the vector $(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_W)$.

The homogeneous association statistic is meant to detect departures from a constant subpopulation-specific effect.

To evaluate the treatment effect in relative scale, we suggest the use of T_1^* and T_3^*, which are the statistics T_1 and T_3 applied to the log of the effect estimates.

4. A supremum statistic on the relative scale:

$$T_1^* = max_{j=1,\cdots,K} |log(\hat{\theta}_j) - log(\hat{\theta}_{ALL})| / \hat{\sigma}_j \tag{3.5}$$

where $\hat{\sigma}_j = \sqrt{\hat{var}(log(\hat{\theta}_j) - log(\hat{\theta}_{ALL}))}$ with $\hat{\theta}_{ALL}$ as the treatment effect for all subjects.

5. A homogeneous association statistic on the relative scale:

$$T_3^* = (log(\hat{\boldsymbol{\theta}}) - log(\hat{\boldsymbol{\theta}}_W))^T \hat{\boldsymbol{\Sigma}}^{-1} (log(\hat{\boldsymbol{\theta}}) - log(\hat{\boldsymbol{\theta}}_W)) \tag{3.6}$$

where $\hat{\boldsymbol{\theta}}$ is the $K-$dimensional vector of treatment effect estimates for the K subpopulations; $\hat{\boldsymbol{\theta}}_W$ is a $K-$dimensional vector with all elements equal to a weighted average of the estimated treatment effects $\hat{\theta}_j$s. Here too, we pick weights $1/\hat{\sigma}_j$, where $\hat{\sigma}_j$ is now the

estimated standard deviation of $log(\hat{\theta}_j)$. $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix of the vector $(log(\hat{\boldsymbol{\theta}}) - log(\hat{\boldsymbol{\theta}}_W))$.

The distribution of T_1, T_2, T_3, T_1^*, and T_3^* can be estimated by sampling repeatedly from the joint asymptotic distribution of $(\hat{\theta}_1, \cdots, \hat{\theta}_K, \hat{\theta}_{ALL})$. Following the approach taken in Bonetti, Zahrieh, Cole and Gelber, 2009, our software implements the permutation-based inference by permuting the covariate values across the patients within each treatment group and then re-computes the test statistics based on the permuted samples. The variances are estimated from the permuted samples. The permutation *p*-value for a particular statistic is the percentage of the times that the permutation based statistic is more extreme than the statistic computed on the observed outcome under the general null hypothesis of no covariate effect.

Note that examination of subpopulation treatment effect patterns may not be possible if the sample size of the trial is insufficient to support such investigations. For GLMs, if the sample size is too small there may be computational problems for estimating parameters, as the fitting algorithm may not converge. Based on our experience, a requirement of at least 10 independent samples for each parameter in the model for each subpopulation is suggested.

3.4 Results

3.4.1 A Simulation Study

We explore the properties of the GLM implementation of STEPP by a simulation study. Specifically, we evaluate the accuracy of Type I error rate α for tests based on the statistics described in Section 3.3. We have not explored the power of these statistics as there are too many possible but yet, no representative alternate scenarios.

Under the null hypothesis, there is no treatment effect heterogeneity across the subpop-

ulations of interest, Z. We generate Z according to a $N(25, 10^2)$ distribution for all our tests. For each of the GLM models, the outcomes are sampled from three different distributions. For the Gaussian model, they are $N(55, 49)$, $N(75, 25)$, and $N(95, 36)$; for the binomial model, they are $Bernoulli(0.3)$, $Bernoulli(0.5)$ and $Bernoulli(0.7)$; and for the Poisson model, they are $Poisson(5)$, $Poisson(10)$ and $Poisson(15)$. Patients are randomly assigned to either treatment arm with probability 0.5.

We generate 500 sample data sets. For each data set, we generate sample data of size n = 100, 200, 500, and 1000. The subpopulations are constructed using sliding window parameters $(r1, r2) = (30, 40), (60, 80), (150, 200),$ and $(300, 400)$.

The *p*-value of each of the sample data set is computed. The estimated α level is percentage of *p*-values below the specified α level. Table 3.2 shows the simulation results for the permutation test with Gaussian model, ($Z \sim N(95, 49)$) case, and, in particular, the estimated α level, with different sample sizes and different sliding window parameter values. Other results (the other two Gaussian models, the binomial models and the Poisson models) are not reported on here for brevity, but the results are similar. The other results can be found in Table A.5 through Table A.12 in Appendix A.3.3.

Some of the permutation tests are slightly inflated in their Type I error probability (α). As the sample size n increases, as one would expect, the results shrink towards the nominal Type 1 error rate.

3.4.2 Analysis of the Aspirin/Folate Polyp Prevention Study Data

The two treatment groups are placebo and 81 mg daily dosage of aspirin. We choose to model the risk, (p), using logistic regression with the outcome being any occurrences of AD. The GLM model within each subpopulation can be written as

$$logit(p) = \beta_0 + \beta_1 \times trt \tag{3.7}$$

The covariate of interest is age, which is treated as a continuous variable. The STEPP subpopulations are created by setting $r2$ to be 100 and $r1$ to be 30. Based on this setting, 8 subpopulations are generated. The subpopulations summary can be found in Table 3.3.

Based on these 8 subpopulations, treatment effect estimates of subpopulations are computed and the resulting STEPP plots are generated (see Figure 3.1, Figure 3.2 and Figure 3.3). Figure 3.1 plots the risk of experiencing AD for the two treatment groups across different age subgroups. It shows that the risk for the placebo group to be higher than the risk for the treatment group for age subgroups in the middle. Figure 3.2 plots the absolute risk difference between the two treatments across different age subgroups. Figure 3.3 plots the odds ratio between the two treatments across different age subgroups. The supremum statistic is displayed in all three plots. The result is highly significant indicating that sampling variability cannot account for the observed patterns.

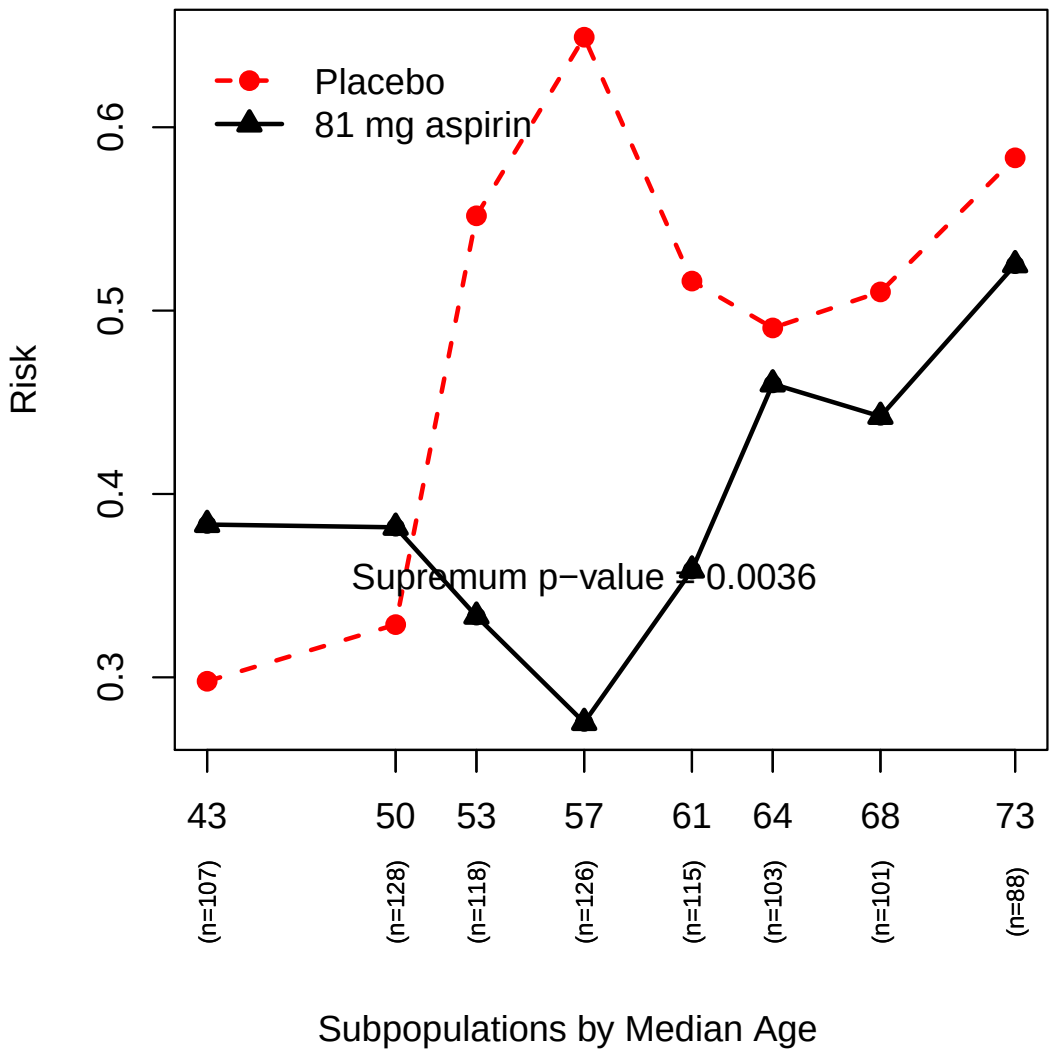

Figure 3.1: The STEPP plot shows the absolute risk (or probability of experiencing AD) for two treatment groups across different age subgroups - the ″red″ dashed line is the placebo group and the ″black″ solid line is the 81 mg aspirin group.

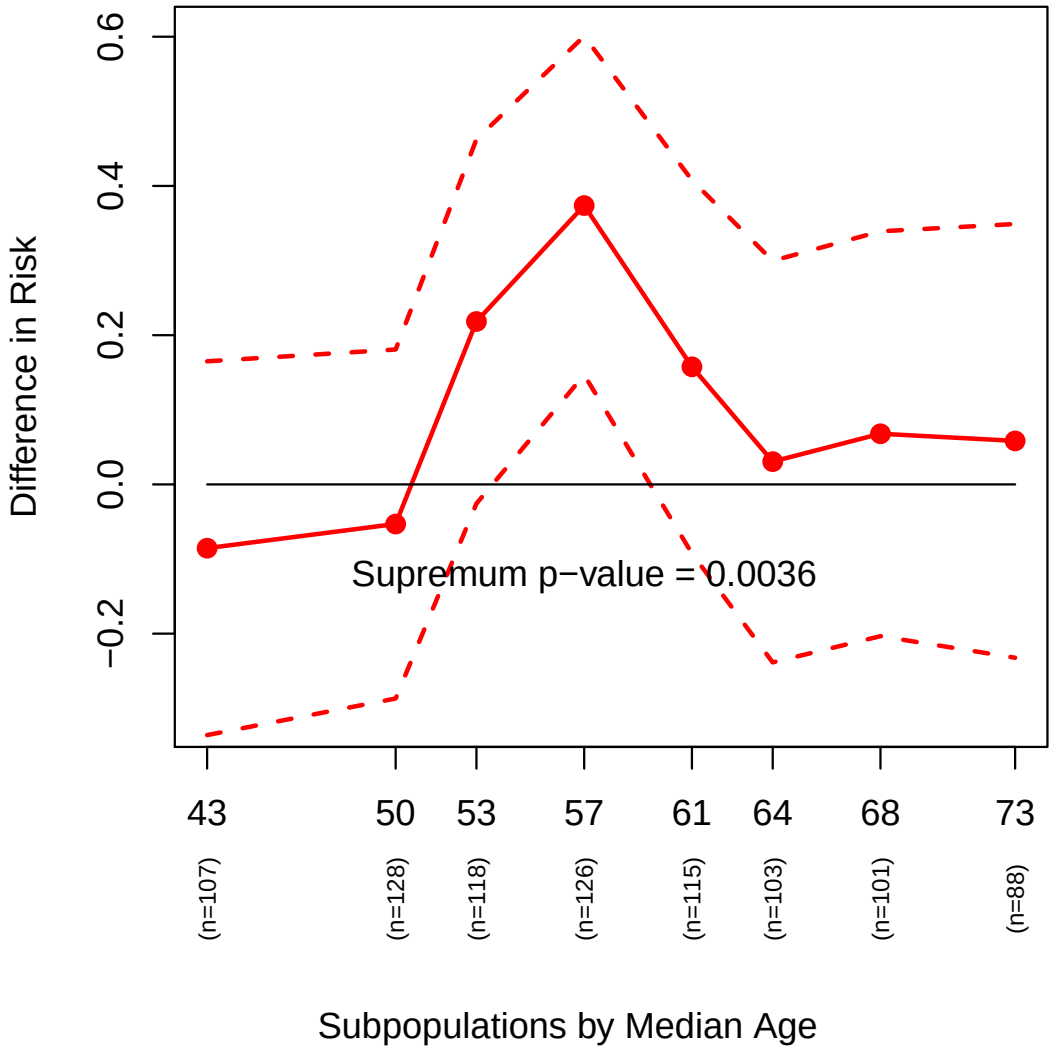

Figure 3.2: The STEPP plot shows the differences in risk of experiencing AD across the various age subgroups between placebo and the 81 mg aspirin treatment group. The interaction supremum p-value based on risk difference is 0.0036, suggesting an interaction effect between the risks and the age-defined subpopulations on the absolute scale. The effect of the 81 mg in reducing the risk of AD compared with placebo appears to be larger for patients in the middle age subpopulations than it is for the youngest and the oldest subpopulations.

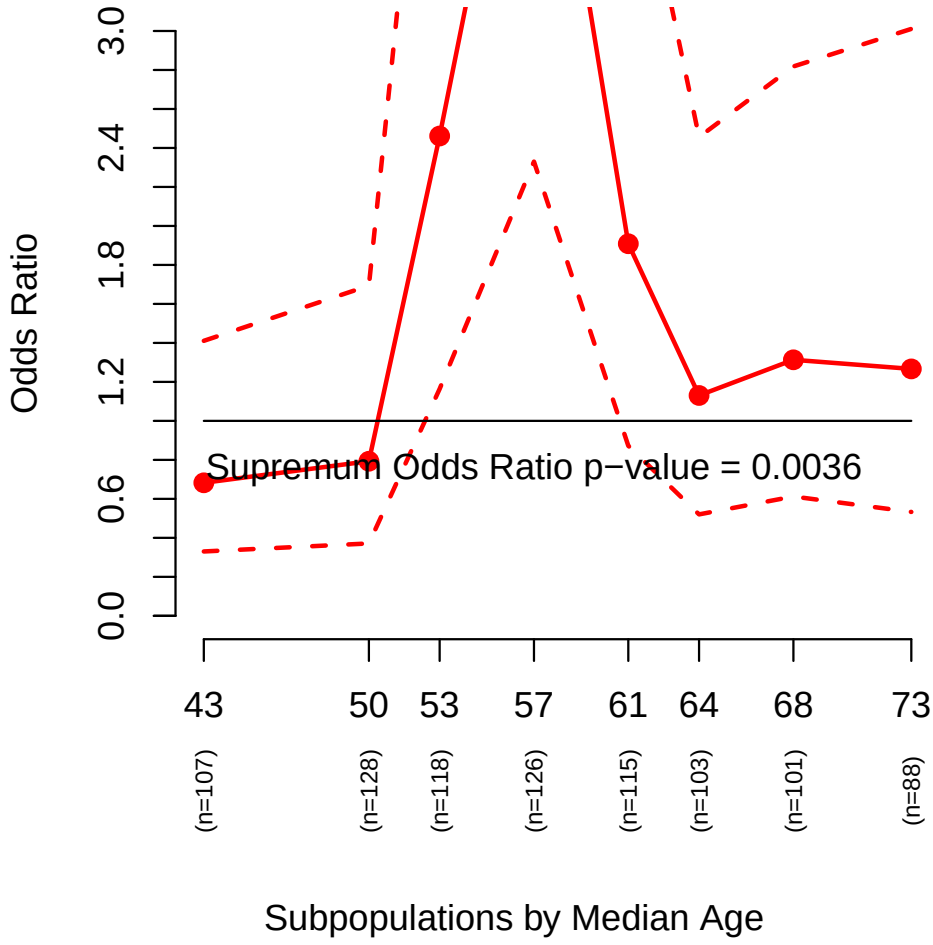

Figure 3.3: The STEPP plot shows the relative risk of experiencing AD across the age subgroups between placebo and the 81 mg aspirin treatment group. The overall odds ratio of experiencing AD is about 1.46 when comparing the two groups. The interaction supremum p-value based on odds ratio is 0.0036, also suggesting a possible interaction effect between risks and the age-defined subpopulations on the relative scale.

3.5 Discussion

STEPP is a graphical tool that assists researchers in exploring the heterogeneity of treatment effects according to the values of a continuous baseline covariate across overlapping subpopulations. From the STEPP plots, one can discern treatment effect differences visually. The p-value that is shown together with each of the plots allows an assessment of the significance of the interaction. Note that the very definition (and therefore the presence) of an interaction depends on both the model and the scale of measurement of the treatment effect, so that a careful exploration of the appropriate metric for the problem at hand should be conducted, and the results interpreted accordingly. Furthermore, as STEPP allows a patient to belong to two or more overlapping subpopulations, the estimation of treatment effects borrows strength from adjacent subpopulations.

As an example, the results of the Aspirin/Folate Polyp Prevention Study data confirm the studies' original result, which shows a moderate beneficial effect on (not) experiencing AD with a daily dosage of 81 mg of aspirin. In addition, it shows graphically that the benefit increases substantially for patients with ages between 50 and 60. The permutation p-value indicates that the interaction is significant. However, as shown in the complete analyses included in the appendices, the benefit diminishes when the daily dosage of 325 mg of aspirin is used. The permutation p-value indicates that any interaction at this dosage may indeed be due to sampling variability. Lastly, it suggests slight risk improvements when comparing the 81 mg daily and 325 mg daily aspirin dosage. The permutation p-value, however, is not significant, suggesting that the interaction effect may also be due to sampling variability. Thus, STEPP not only confirms the original findings, but also points to the age subgroup which may benefit the most from taking low-dose aspirin.

STEPP is non-parametric in nature with respect to the interaction effect, and it allows one to examine possible complex interaction effects. For the sliding window, one can adjust the two window parameters to explore potential different interaction patterns. As a consequence, the p-value obtained from any STEPP analysis does depend upon the

specific choice of the parameter values (e.g., the two parameters for the sliding window approach) and as such it should be interpreted with caution. It is recommended that several different smoothing parameters be investigated in sensitivity analyses to assess the robustness of the results. One must balance the granularity of subpopulation treatment effects, the larger variability of estimates obtained from sparse subpopulations sizes, and the smoothness of the STEPP plots.

The current work and software have some limitations: (i) it is restricted to continuous, binary and count data modeled by standard GLMs; (ii) it restricts the analysis to the comparison of two treatment groups; (iii) it allows the examination of the interaction effect with only one covariate of interest. However, one may also use as the covariate of interest a baseline risk score that can be a function of several individual patient and disease characteristics, as was done in Viale et al., 2008. Another approach is to use principal component analysis (PCA) to generate the principal components based on these characteristics, and then use the principal components as the covariate of interest in STEPP, as was done in Pogue-Geile, 2013.

Currently, the p-value of the statistics are being provided using a permutation approach which are computational intensive. The Type I error rate are slightly inflated but as the sample size increases, the results shrink towards the nominal rate. A marginal model approach has been suggested. We are evaluating this approach and will make a comparison with the permutation approach in a future research effort.

More in general, one still needs to adjust for multiple testing if several different covariates are evaluated one at a time. In addition, the approach does not address the issue of post-hoc analysis as opposed to pre-specified analysis, as well as issues of confounding if the analysis is based on retrospective exposure assignments as opposed to randomized trials. As is the case with any exploration of subgroup treatment effects, hypothesis generating analyses should be distinguished from those intended to evaluate pre-specified hypothesis.

Overall, it should be noted that STEPP is an exploratory tool. In particular, it is not meant to be used to determine specific cut-points in the range of values of the covariate of interest, but rather to provide some indication regarding the ranges of values where the treatment effect might have a particularly beneficial (or detrimental) effect. The permutation p-value indicating the statistical significance of the observed heterogeneity should always be presented together with the graphical presentation of STEPP to avoid over-interpretation of the graphical result. Also, results should be confirmed using results from other data sets investigating similar treatment comparisons. Future research work is needed to investigate how to use STEPP to identify cut-points based on cross validation studies (Pogue-Geile, 2013).

Now that STEPP has been extended to handle models for outcomes other than survival, it will enhance the investigators' toolset to explore heterogeneity of treatment effects based on these models. It provides visual information for hypothesis generation purposes, and it can inform clinical decision-making in the direction of customizing patient care.

3.6 Software

In a concerted effort, the existing **stepp** R package was updated to handle the three GLM models. The latest version of **stepp** (version 3.0.10) is available through CRAN together with R version 3.1.1. The interface for STEPP analysis is completely redesigned so that all different statistical models can be specified and analyzed consistently. The new interface is implemented using S4 objects. However, the old STEPP functions still work but will be deprecated iin the future.

The following is a summary of the new interface:
`stepp.win` - create a stepp window (`stwin`) object with r1 and r2 as parameters;
`stepp.subpop` - create a steppo subpopulation (`stsubpop`) object. The `generate`

method is used to fill in the subpopulation;
`stepp.CI, stepp.KM, stepp.COX, stepp.GLM` - constructor functions to create the corresponding S4 stepp models: `stmodelCI, stmodelKM, stmodelCOX and stmodelGLM`. The `summary, print` and `plot` methods of each of the model generate the resulting tables and the three stepp plots for analysis.
`stepp.estimate` - apply the stepp model to the subpopulations and estimate their effects; and
`stepp.test` - apply permutation test to detect the null hypothesis of no heterogeneity among the subpopulations and to produce all the different estimates, variance covariance matrices and pvalues from various statistics.

Two pseudo data sets are provided (we cannot post the original data set, so we post modified versions):
`aspirin` - aspirin study by John Baron et al
`big` - big breast cancer study

For backward compatibility with previous versions of STEPP, the old interfaces are maintained:
`analyze.CumInc.stepp, analyze.KM.stepp, stepp, stepp_summary, stepp_print, stepp_plot`.

The reference manual available together with the software on CRAN contains the details about these functions and S4 objects.

Acknowledgments

The Aspirin/Folate Polyp Prevention Study Group collaborative is acknowledged for permission to use the data from the Aspirin/Folate Prevention as an illustrative example. Partial support was provided by Grant No. CA-075362 from the United States National Cancer Institute, Grant No. P30 DE020752, and grant 2007AYHZWC from the Italian Ministry of Education, University and Research. We would also like to acknowledge the support to Wai-Ki Yip by the Clinical Epidemiology of Lung Diseases Training Grant (T32 HL007427) and (2 OF 2) GENETIC EPIDEMIOLOGY OF COPD Grant (5R01HL089856-08). We also like to thank the following individuals for many insightful discussions: Prof Christoph Lange, Prof Nan Laird, Christina McIntosh, Licette CY Liu, and Matteo Puntoni.

Table 3.1: Treatment effects of the GLM models. The model $E(Y|trt, X) = trt \times \alpha + X\beta$ is fitted for the entire population and each subpopulation. X denotes any covariates (excluding the treatment indicator) of the subpopulation. $\overline{X}$ denotes the mean values of those covariates.

GLM model	Treatment Outcome	Treatment Effect in Absolute Scale	Treatment Effect in Relative Scale
Gaussian $E(Y\|trt, X) = trt \times \alpha + X\beta$	Expected outcome of the model, $E(Y\|trt, \overline{X})$	Difference of expected outcomes between the two treatment arms, $E(Y\|trt = 1, \overline{X}) - E(Y\|trt = 0, \overline{X})$	Ratio of expected outcomes between the two treatment arms $E(Y\|trt = 1, \overline{X})/ E(Y\|trt = 0, \overline{X})$
Binomial $logit(E(Y\|trt, X)) = trt \times \alpha + X\beta$	Expected outcome of the model, $E(Y\|trt, \overline{X}) = p$, is the risk	Difference of risks between the two treatment arms, $p_{trt=1,\overline{X}} - p_{trt=0,\overline{X}}$	Ratio of risks between the two treatment arms $p_{trt=1,\overline{X}}/p_{trt=0,\overline{X}}$
Poisson $log(E(Y\|trt, X)) = trt \times \alpha + X\beta$	Expected outcome of the model, $E(Y\|trt, \overline{X}) = \lambda$, is the rate	Difference of rates between the two treatment arms, $\lambda_{trt=1,\overline{X}} - \lambda_{trt=0,\overline{X}}$	Ratio of rates between the two treatment arms $\lambda_{trt=1,,\overline{X}}/\lambda_{trt=0,,\overline{X}}$

Table 3.2: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Gaussian model N(95,36). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2. Values in *italics* indicate the cases when the results appear to be anti-conservative.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.012	0.046	0.096
			T2	0.008	0.046	0.096
			T3	0.010	0.050	0.094
			T1*	0.012	0.046	0.094
			T3*	0.012	*0.032*	0.056
200	60	80	T1	0.010	*0.062*	0.116
			T2	0.010	0.042	0.098
			T3	0.016	0.046	0.094
			T1*	0.010	*0.066*	0.114
			T3*	0.016	0.042	0.096
500	150	200	T1	0.008	*0.064*	0.108
			T2	0.012	0.052	0.094
			T3	0.010	0.042	0.082
			T1*	0.008	*0.066*	0.106
			T3*	0.010	0.042	0.084
1000	300	400	T1	0.012	0.048	0.094
			T2	0.008	0.050	0.106
			T3	0.008	0.050	0.104
			T1*	0.012	0.046	0.092
			T3*	0.008	0.050	0.102

Table 3.3: The subpopulation summary for the Aspirin/Folate Polyp Prevention Study Data using age as the covariate of interest. The number of patients per subpopulation (r2) is 100 and the largest number of patients in common among consecutive subpopulations (r1) is 30. The number of subpopulations created is equal to 8.

Subpopulation	Size	Median Age	Minimum Age	Maximum Age
1	107	43	29	47
2	128	50	46	51
3	118	53	52	55
4	126	57	55	59
5	115	61	59	62
6	103	64	62	66
7	101	68	66	71
8	88	73	70	78

References

Barcella, W. (2013). Comparison between the permutation-based and the GEE-based approach to STEPP for the analysis of the subpopulation treatment effects in clinical trials. *Bocconi University, MSc Thesis*.

Baron, J. A. and others, R. J. (2003). A Randomized Trial of Aspirin to Prevent Colorectal Adenomas. *The New England Journal of Medicine* **348,** 891–899.

Bonetti, M. and Gelber, R. (2000). A Graphical Method to Assess Treatment-Covariate Interactions Using the Cox Model on Subsets of the Data. *Statistics in Medicine* **19,** 2595–2609.

Bonetti, M. and Gelber, R. (2004). Patterns of Treatment Effects in Subsets of Patients in Clinical Trials. *Biostatistics* **53,** 465–481.

Bonetti, M., Zahrieh, D., Cole, D., and Gelber, R. (2009). A Small Sample Study of the STEPP Approach to Assessing Treatment-Covariate Interactions in Survival Data. *Statistics in Medicine* **28(8),** 1255–1268.

Colleoni, M. and others (2002). Duration of Adjuvant Chemotherapy for Breast Cancer: a Joint Analysis of Two Randomized Trials Investigating Three Versus Six Course of CMF. *British Journal of Cancer* **86,** 1705–1714.

Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society* **34,** 187–220.

Fine, J. and Gray, R. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* **94,** 496–509.

Gray, R. (1998). A Class of K-sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of Statistics* **16,** 1141–1154.

Lagakos, S (2006). The Challenge of Subgroup Analysis - Reporting Without Distorting. *The New England Journal of Medicine* **354,** 1667–1669.

Lazar, A., Cole, B., Bonetti, M. and Gelber, R (2010). Evaluation of Treatment-Effect Heterogeneity Using Biomarkers Measured on a Continuous Scale: Subpopulation Treatment Effect Pattern Plot. *Journal of Clinical Oncology* **28(29),** 4539–4544.

Pesarin, F. (2001). Multivariate permutation tests: with application in Biostatistics. *Wiley, New York*.

Pocock, S. (2008). More on Subgroup Analysis in Clinical Trials. *The New England Journal of Medicine* **358,** 2076–2077.

Pogue-Geile, K. L. and others (2013). Predicting Degree of Benefit from Adjuvant Trastuzumab in NSABP Trial B-31. *Journal of National Cancer Institute* **105(23),** 1782–1788.

Viale, G. and others (2008). Prognostic and Predictive Value of Centrally Reviewed Ki-67 Labeling Index in Postmenopausal Women with Endocrine-Responsive Breast Cancer: Results from Breast International Group Trial 1-98 Comparing Adjuvant Tamoxifen with Letrozole. *Journal of Clinical Oncology* **28(34),** 5569–5575.

Wang, R., Lagakos, S. and Ware, J. (2007). Statistics in Medicine - Reporting of Subgroup Analysis in Clinical Trials. *New England Journal of Medicine* **357,** 2189–2194.

R Foundation for Statistical Computing, Vienna, Austra. ISBN 3-90051-07-0 (2008). R: A Language and Environment for Statistical Computing. *http://CRAN.R-project.org*.

Yip, W.-K. (2011). *stepp: Subpopulation Treatment Effect Pattern Plot (STEPP): R package version2.3-2 http://CRAN.R-project.org/package=stepp*.

A. Appendices

A.1 A Novel Method for Detecting Association Between DNA Methylation and Diseases Using Spatial Information

A.1.1 Additional Application Results: Application of the SCM to chromosome 10 of a cancer dataset

Similar to what we did to chromosome 14, we scanned the entire chromosome 10 from the beginning to the end with K, the number of consecutive CpG sites for the sliding window size, as 51. The scanning result showed that there were 3726 windows with p-values $< 10^{-5}$, which is the Bonferroni corrected significance level. Many of these significant windows were contiguous forming 106 clusters of statistically significant regions. (Two windows belong to a window cluster if they have 4 or less non-significant windows in betwee.) The largest 3 window clusters comprised of 196, 222 and 291 windows. These three window clusters are shown as "'red"' vertical lines in figure A.1.

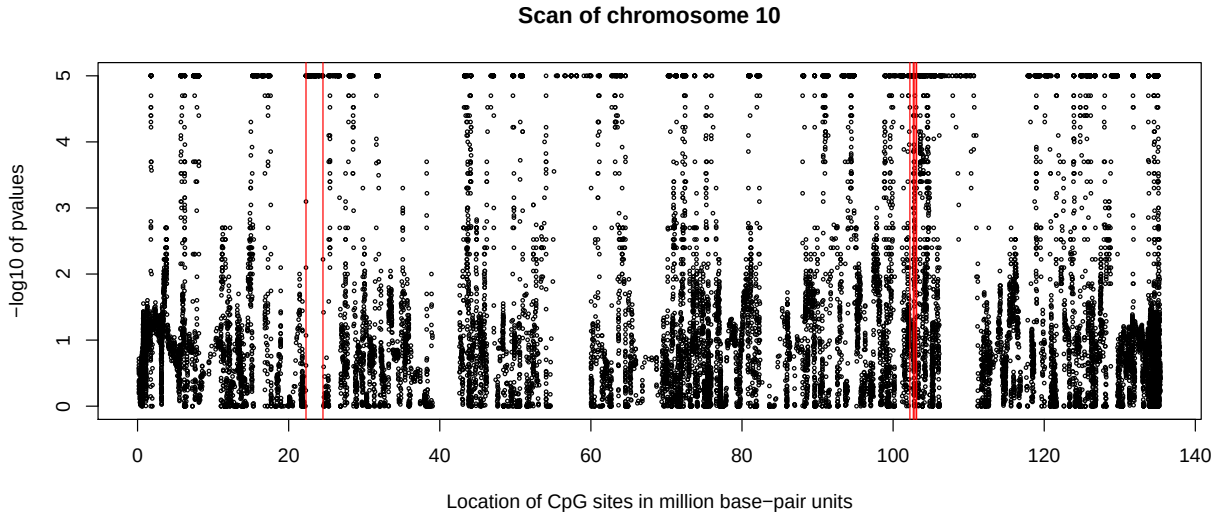

Figure A.1: Using windows of 51 CpG sites, the sliding window scan shows regions of chromosome 10 which have p-values of $< 10^{-5}$ by using the SCM.

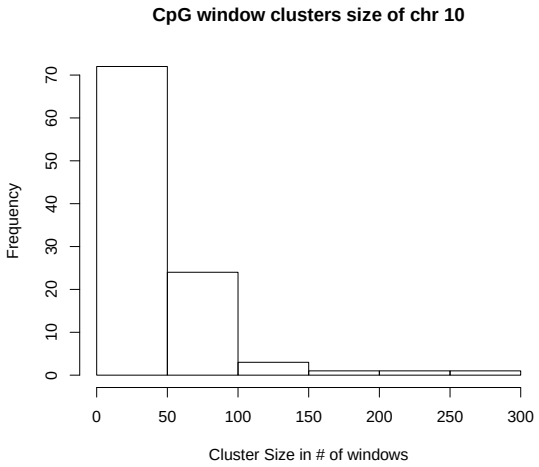

Figure A.2: Histogram of the size of CpG window clusters on chromosome 10 with p-values $< 10^{-5}$.

Table A.1: SABRE: The number of sites and the % of sites selected for each δ.

δ	no. of sites	% of sites
0.00	450725	100.00
0.05	231094	51.27
0.10	137031	30.40
0.15	77794	17.26
0.20	42971	9.53
0.25	24740	5.49
0.30	14988	3.33
0.35	9608	2.13
0.40	6539	1.45
0.45	4584	1.02
0.50	3320	0.74
0.55	2356	0.52
0.60	1557	0.35
0.65	937	0.21
0.70	524	0.12
0.75	222	0.05
0.80	69	0.02
0.85	17	$<$0.01
0.90	2	$<$0.01

A.2 A principal component approach for the detection of unknown substructure in DNA methylation data

A.2.1 Complete SABRE cohort analysis

The purpose of the investigation was to explore the population substructure in the selected subset of the SABRE cohort data set.

We first applied Mclust to cluster all the CpG sites using a mixture of two Gaussian models. Then, the differences of the mean of these Gaussian models were computed. Table A.1 summarizes the result.

For each set K_δ, the methylation distance matrix was created and the principal coordinate

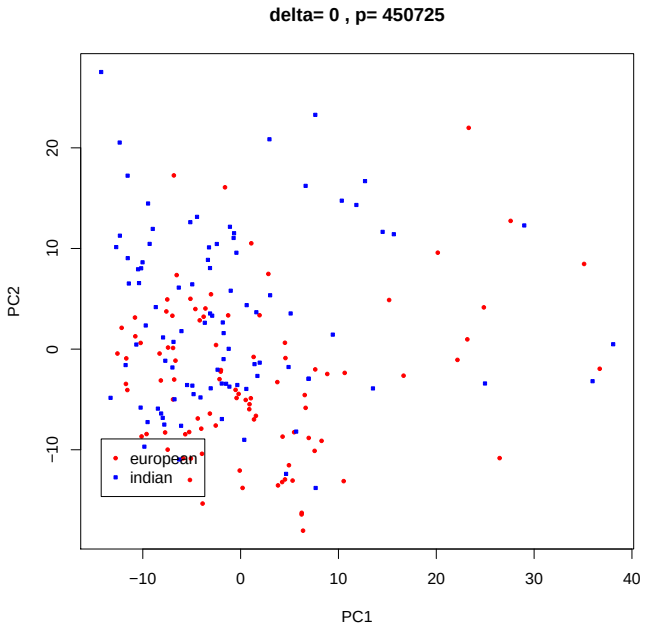

Figure A.3: SABRE: Using all 450725 methylation sites, i.e. with $\delta \geq 0$, the two subpopulations are mixed together.

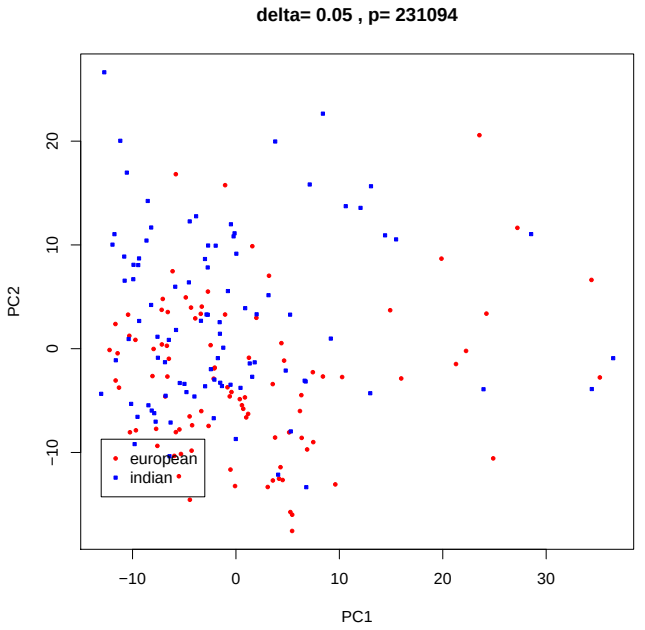

Figure A.4: SABRE: Using 231094 methylation sites, i.e. with $\delta \geq 0.05$, the two subpopulations are mixed together.

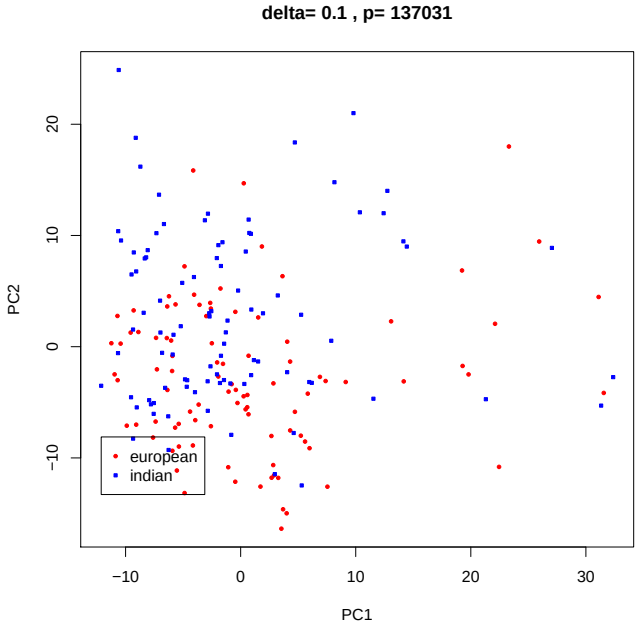

Figure A.5: SABRE: Using 137031 methylation sites, i.e. with $\delta \geq 0.10$, the two subpopulations are mixed together.

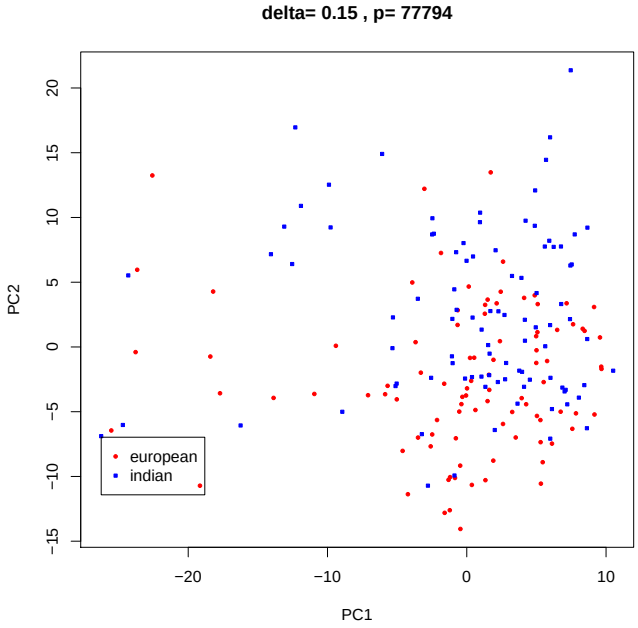

Figure A.6: SABRE: Using 77794 methylation sites, i.e. with $\delta \geq 0.15$, the two subpopulations are mixed together.

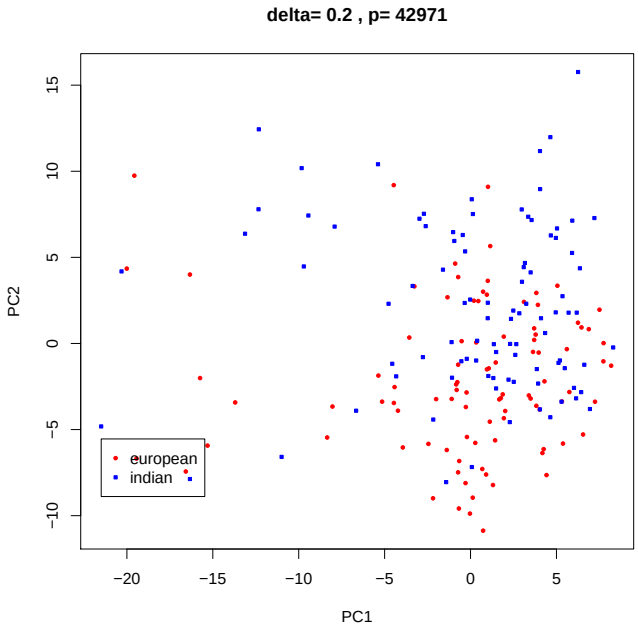

Figure A.7: SABRE: Using 42971 methylation sites, i.e. with $\delta \geq 0.20$, the two subpopulations are mixed together.

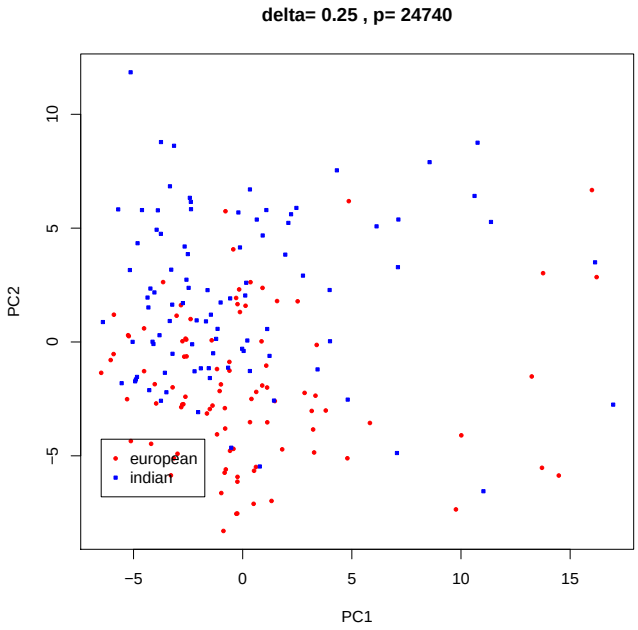

Figure A.8: SABRE: Using 24740 methylation sites, i.e. with $\delta \geq 0.25$, the two subpopulations are mixed together.

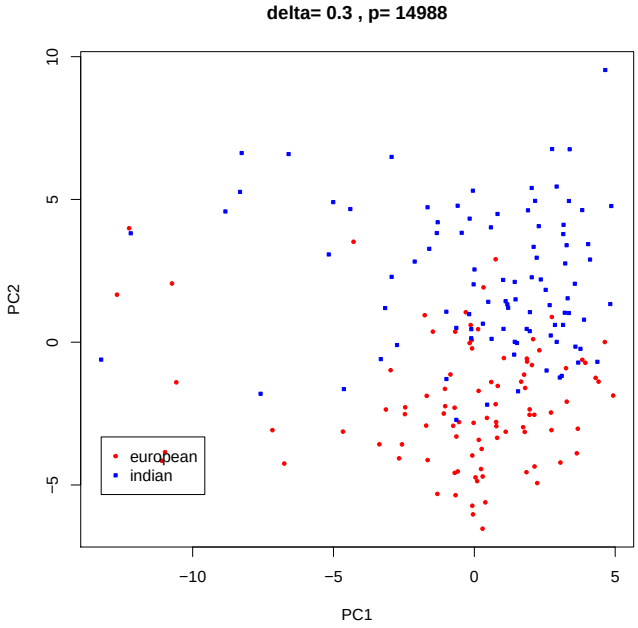

Figure A.9: SABRE: Using 14988 methylation sites, i.e. with $\delta \geq 0.30$, the two subpopulations are mixed together.

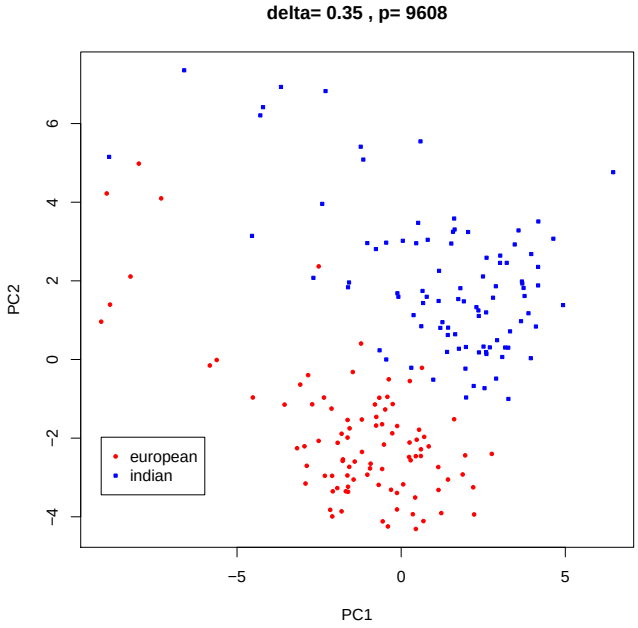

Figure A.10: SABRE: Using 9608 methylation sites, i.e. with $\delta \geq 0.35$, the two subpopulations are mixed together.

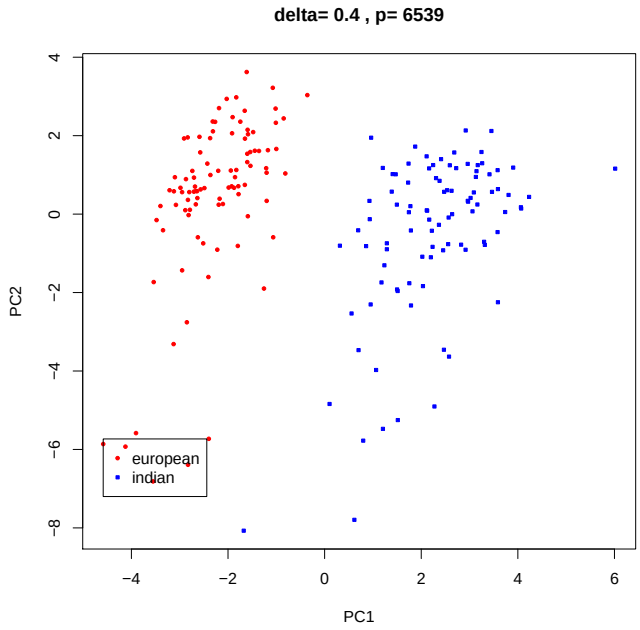

Figure A.11: SABRE: With $\delta \geq 0.40$, 6539 methylation sites are identified. The two clusters representing the two ethnic groups are now discernible.

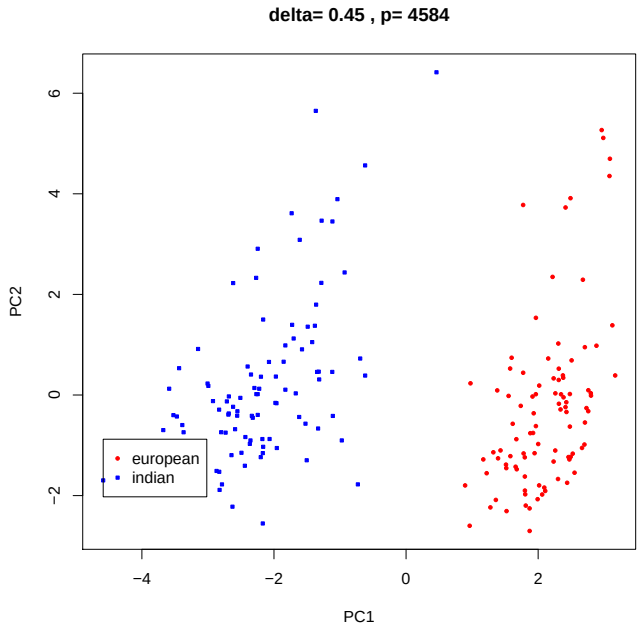

Figure A.12: SABRE: With $\delta \geq 0.45$, 4584 methylation sites are identified. The two clusters representing the two ethnic groups are discernible.

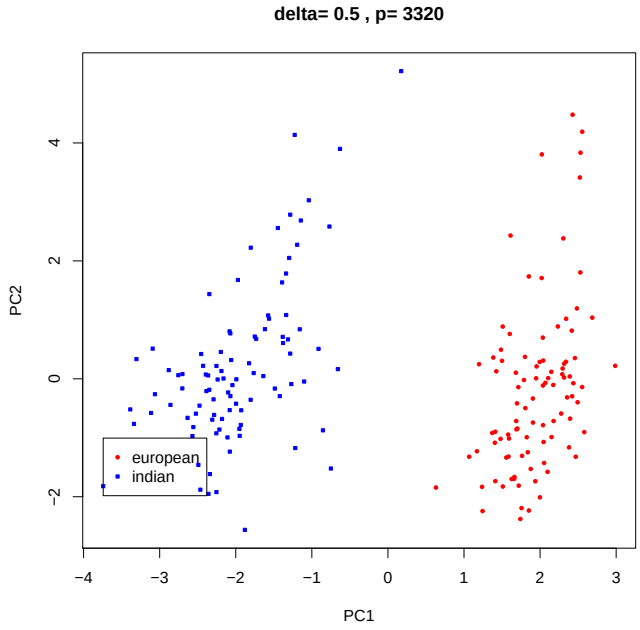

Figure A.13: SABRE: With $\delta \geq 0.50$, 3320 methylation sites are identified. The two clusters representing the two ethnic groups are discernible.

analysis was applied. The following sequence of plots showed the changes as the value of δ was increased. We varied δ from 0 to 0.9 with an increment of 0.05.

A.2.2 % of relevant methylation sites captured by MDA in the SABRE cohort analysis

We did not know which methylation sites were related to the population substructure. However, we knew the identification of the subpopulation for all our samples in the SABRE cohort. So, if we assumed that the substructure was related to the shift in the mean of the methylation β-values for each sites, we could estimate the relevant methylation sites captured by the Gaussian mixture model scheme by calculating the proportion of the methylation sites that had the biggest difference in mean β-values captured by MDA.

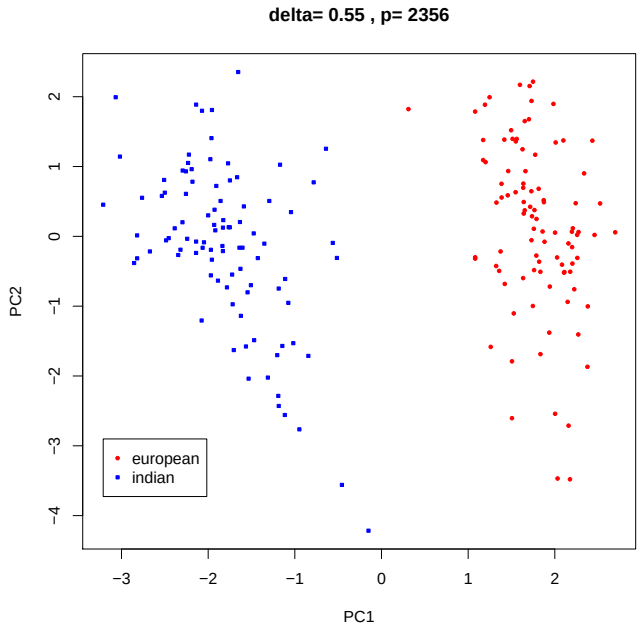

Figure A.14: SABRE: With $\delta \geq 0.55$, 2356 methylation sites are identified. The two clusters representing the two ethnic groups are discernible.

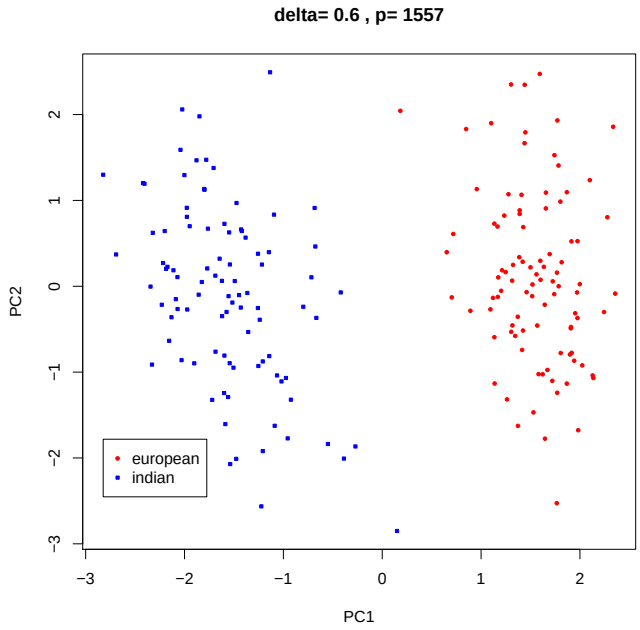

Figure A.15: SABRE: With $\delta \geq 0.60$, 1557 methylation sites are identified. This seems to be the clearest separation.

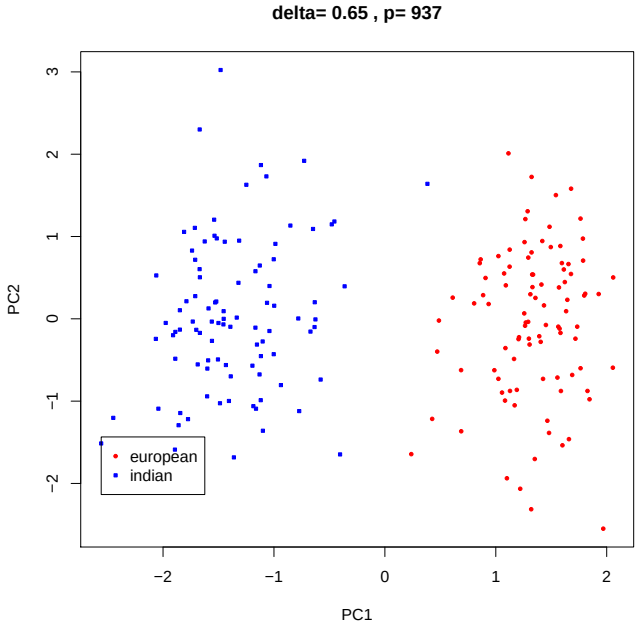

Figure A.16: SABRE: With $\delta \geq 0.65$, 937 methylation sites are identified. The two clusters representing the two ethnic groups are discernible.

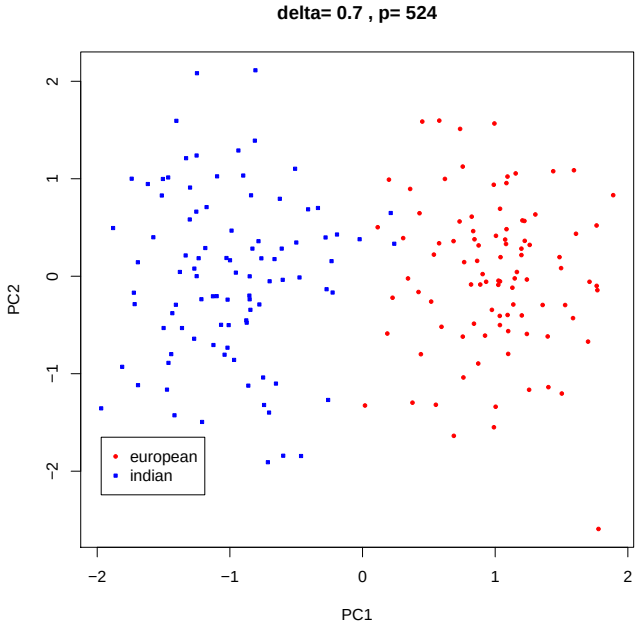

Figure A.17: SABRE: With $\delta \geq 0.70$, 524 methylation sites are identified. The two clusters representing the two ethnic groups are hardly discernible.

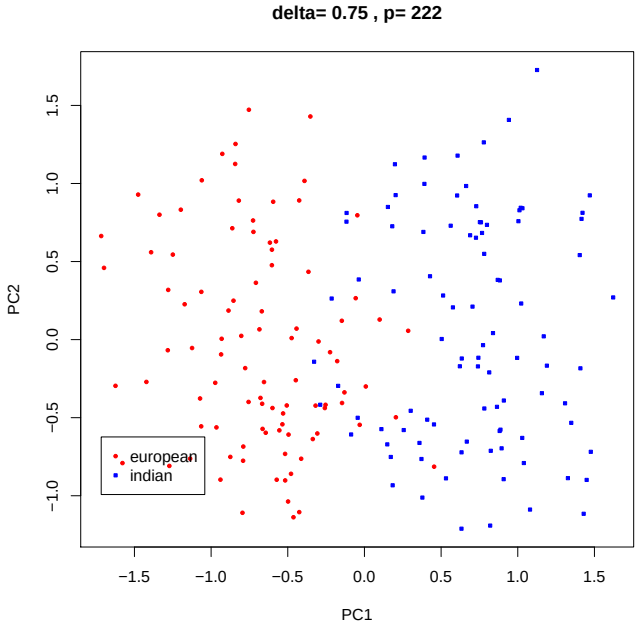

Figure A.18: SABRE: With $\delta \geq 0.75$, 222 methylation sites are identified. The two clusters representing the two ethnic groups are merged into one.

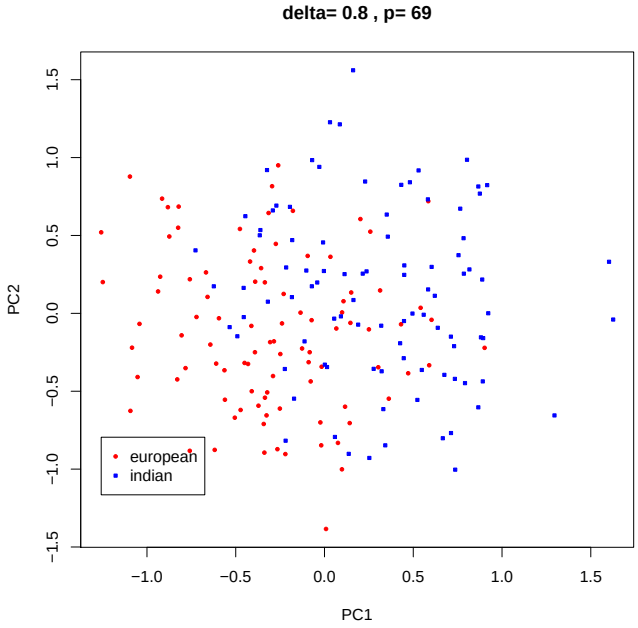

Figure A.19: SABRE: With $\delta \geq 0.80$, 69 methylation sites are identified. The two subpopulations are mixed together.

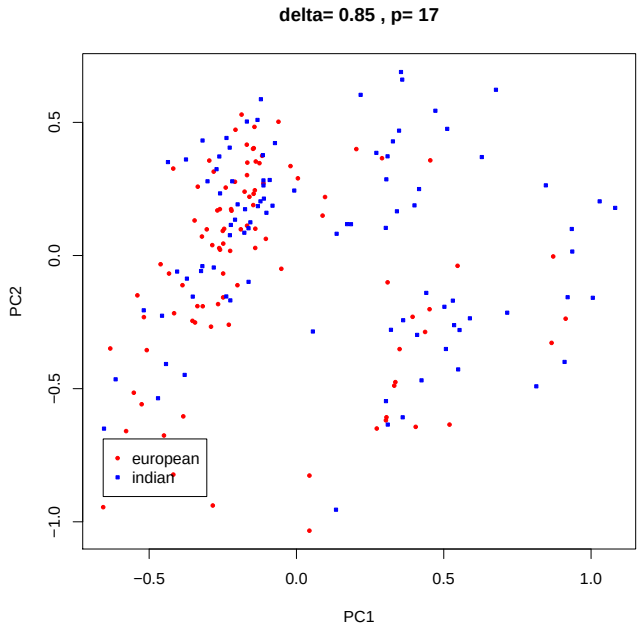

Figure A.20: SABRE: With $\delta \geq 0.85$, only 17 methylation sites are identified. The two clusters are now mixed together again.

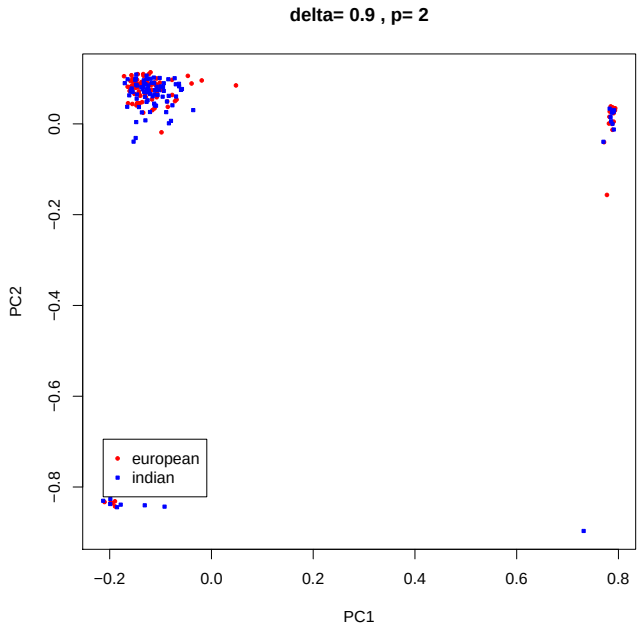

Figure A.21: SABRE: With $\delta \geq 0.9$, 2 methylation sites are identified. There seems to be four clusters here. However, this pattern is not consistent.

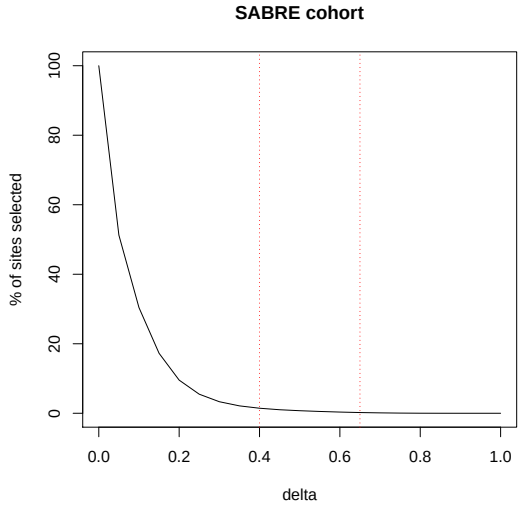

Figure A.22: SABRE: The % of methylation sites selected for the set K_δ is plotted against the specific delta δ for the SABRE cohort. The clustering patterns are observed between the two "red" dotted lines.

Table A.2: SABRE: The columns represent the δs/number of methylation sites for the MDA and the rows represent the δs/number of methylation sites for the difference in mean of the two ethnic subgroups. The result in each cell stands for the number of methylation sites captured/proportion captured by MDA.

	0.85/ 17	0.75/ 222	0.65/ 937	0.55/ 2356	0.45/ 4584	0.40/ 6539	0.35/ 9608	0.25/ 24740	0.15/ 77794
0.30/ 6	0/ 0%	1/ 16.7%	6/ 100%	6/ 100%	6/ 100%	6/ 100%	6/ 100%	6/ 100%	6/ 100%
0.25/ 22	0/ 0%	4/ 18.2%	17/ 77.3%	20/ 90.9%	22/ 100%	22/ 100%	22/ 100%	22/ 100%	22/ 100%
0.21/ 47	0/ 0%	8/ 17%	27/ 57.4%	41/ 87.2%	46/ 97.9%	47/ 100%	47/ 100%	47/ 100%	47/ 100%
0.15/ 234	1/ 0.43%	27/ 11.5%	94/ 40.2%	152/ 65%	195/ 83.3%	214/ 91.5%	225/ 96.2%	233/ 99.6%	234/ 100%
0.08/ 1626	7/ 0.43%	78/ 4.8%	263/ 16.2%	513/ 31.5%	750/ 46.1%	909/ 55.9%	1047/ 64.4%	1394/ 85.7%	1612/ 99.1%
0.05/ 66586	8/ 0.12%	119/ 1.8%	423/ 6.4%	830/ 12.6%	1293/ 19.6%	1627/ 24.7%	2019/ 30.7%	3386/ 51.4%	5610/ 85.2%
0.01/ 110741	16/ <0.1%	193/ 0.17%	789/ 0.71%	1901/ 1.7%	3548/ 3.2%	4940/ 4.5%	7034/ 6.4%	16898/ 15.3%	47549/ 42.9%

As shown in Table A.2, there were overlaps in the table indicating that the Gaussian mixture modeling can pick up relevant CpG sites. In fact, if the difference of mean β-value population subgroups was 0.15, MDA actually picked up a majority of these sites. Although Gaussian mixture modeling did not identify the subgroups, it picked up enough relevant methylation sites so that clustering was possible unless the difference is small (< 0.05). The algorithm was tolerant of noises - up to an additional 6000 non-related CpG sites before the signal disappeared.

A.2.3 Complete colorectal cancer data from TCGA analysis

The purpose of the investigation was to explore the substructure in the TCGA sample data set. Since there was no other information, it would be interesting to see if there were other confounding factors in the data. After data cleaning, we used only 385,885 methylation sites for our analysis here.

We applied Mclust for all CpG sites with a mixture of two Gaussian models. We used the differences in the mean of these Gaussian models for our feature/variable selection.

We first applied Mclust to cluster all the CpG sites using a mixture of two Gaussian models. Then, the differences of the mean of these Gaussian models were computed. Table A.3 summarizes the result.

A.2.4 % of relevant methylation sites captured by MDA in the TCGA data analysis

We did not know which methylation sites were related to cancer. However, we knew the identification of the cancer tissues for all our samples. So, if we assumed that cancer

Table A.3: TCGA: The number of sites and the % of sites selected for each δ.

δ	no. of sites	% of sites
0.00	385885	100.00
0.05	270817	70.18
0.10	223327	57.87
0.15	178294	46.20
0.20	135557	35.13
0.25	94916	24.60
0.30	59954	15.54
0.35	34033	5.82
0.40	16997	4.40
0.45	7510	1.94
0.50	2915	0.76
0.55	1000	0.26
0.60	294	0.08
0.65	59	0.02
0.70	12	<0.01
0.75	1	<0.01
0.80	1	<0.01

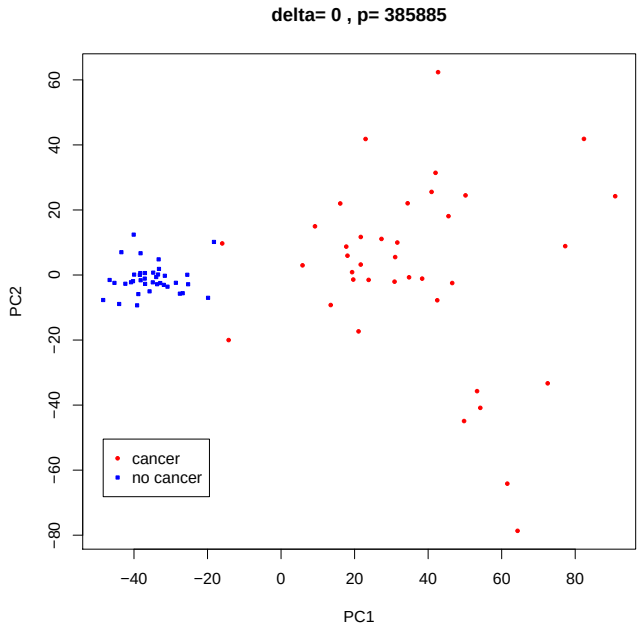

Figure A.23: TCGA: Using all 385885 methylation sites, i.e. with $\delta \geq 0$, it shows two distinctive subgroups: a dense non-cancerous cluster on the left and amore scattered cancerous cluster on the right. This pattern persists for most δs..

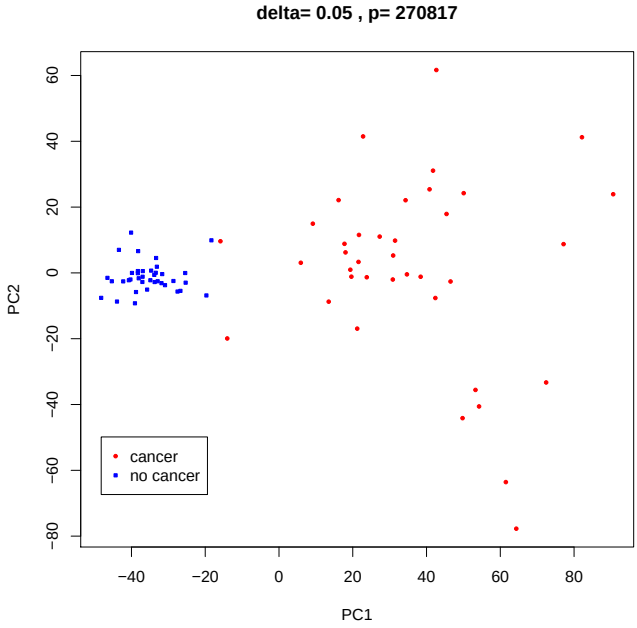

Figure A.24: TCGA: Using 270817 methylation sites, i.e. with $\delta \geq 0.05$, it shows two distinctive subgroups.

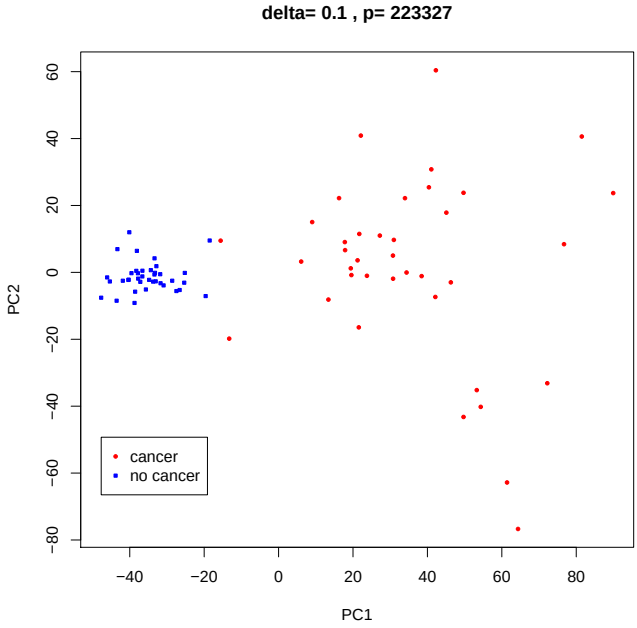

Figure A.25: TCGA: Using 223327 methylation sites, i.e. with $\delta \geq 0.10$, it shows two distinctive subgroups.

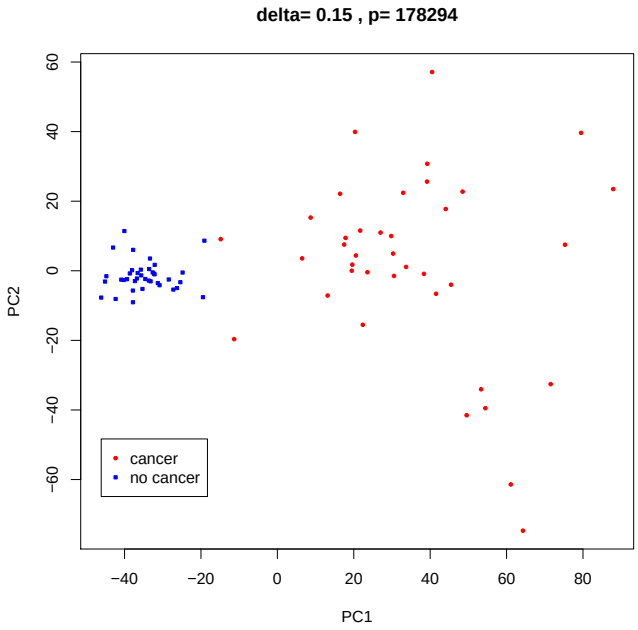

Figure A.26: TCGA: Using 178294 methylation sites, i.e. with $\delta \geq 0.15$, it shows two distinctive subgroups.

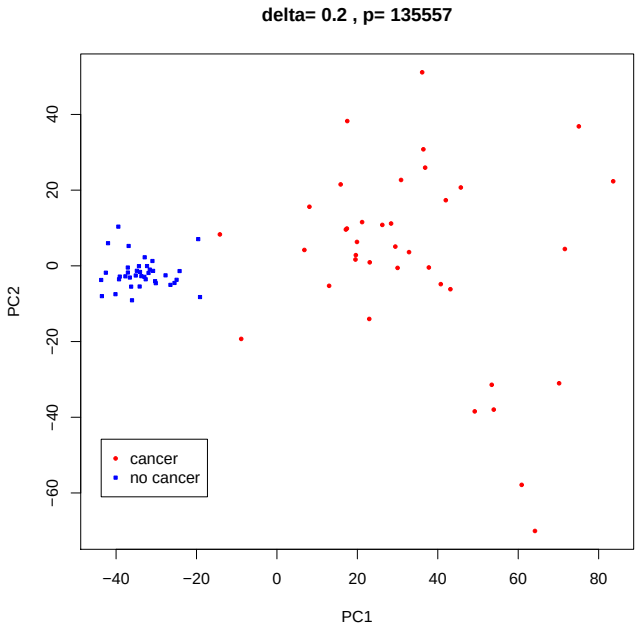

Figure A.27: TCGA: Using 135557 methylation sites, i.e. with $\delta \geq 0.20$, it shows two distinctive subgroups.

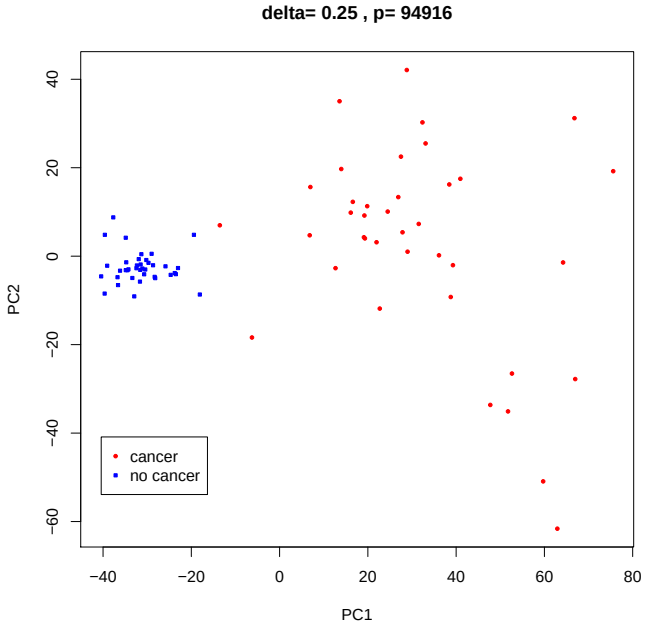

Figure A.28: TCGA: Using 94916 methylation sites, i.e. with $\delta \geq 0.25$, it shows two distinctive subgroups.

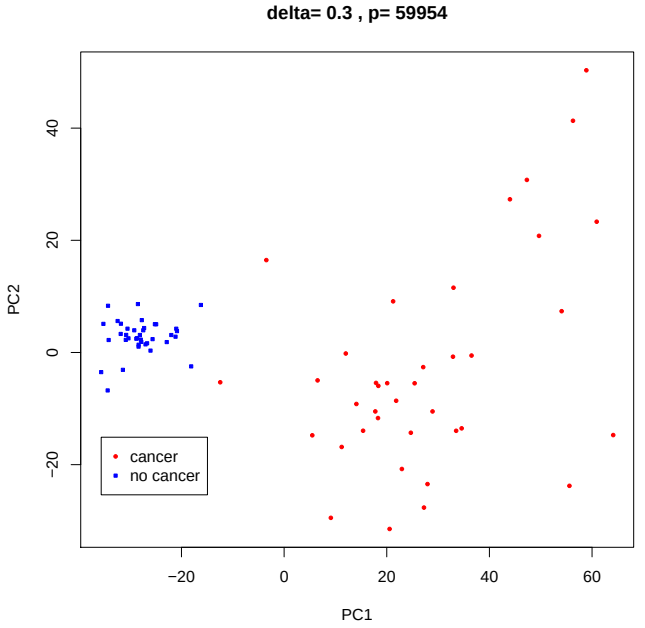

Figure A.29: TCGA: Using 59954 methylation sites, i.e. with $\delta \geq 0.30$, it shows two distinctive subgroups.

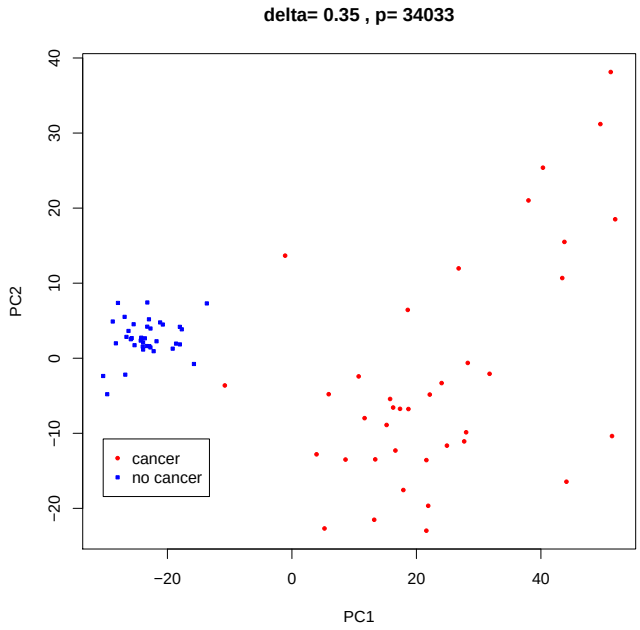

Figure A.30: TCGA: Using 34033 methylation sites, i.e. with $\delta \geq 0.35$, it shows two distinctive subgroups.

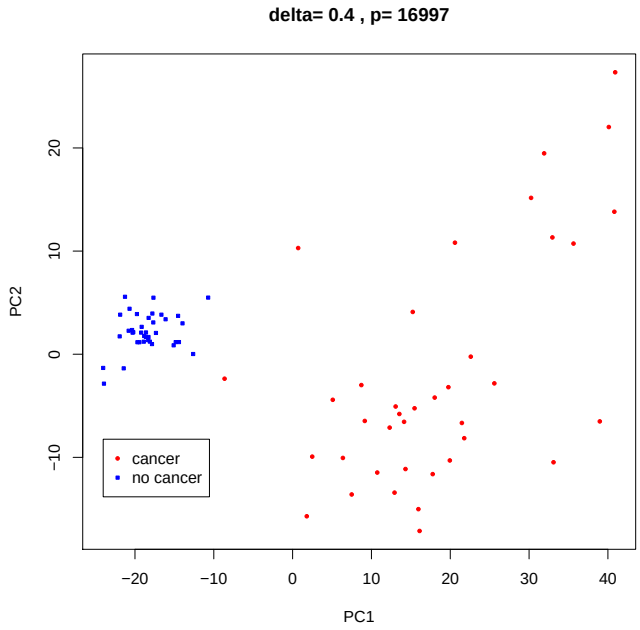

Figure A.31: TCGA: Using 16997 methylation sites, i.e. with $\delta \geq 0.40$, it shows two distinctive subgroups

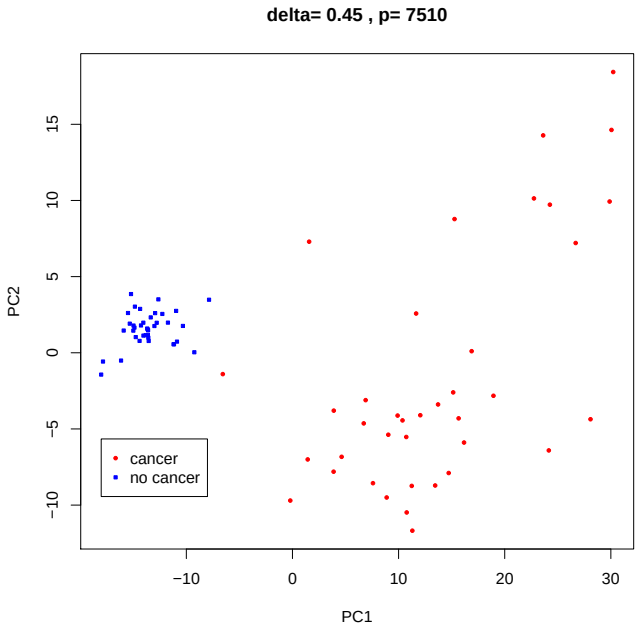

Figure A.32: TCGA: Using 7510 methylation sites, i.e. with $\delta \geq 0.45$, it shows two distinctive subgroups.

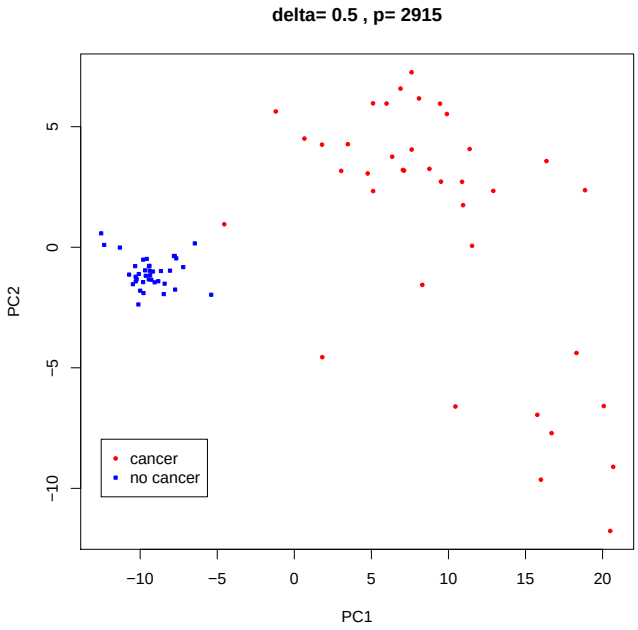

Figure A.33: TCGA: Using 2915 methylation sites, i.e. with $\delta \geq 0.50$, it shows two distinctive subgroups.

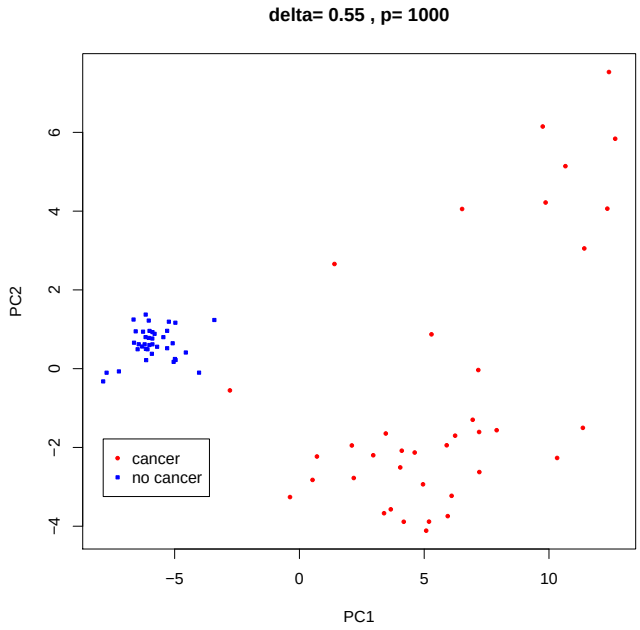

Figure A.34: TCGA: Using 1000 methylation sites, i.e. with $\delta \geq 0.55$, it shows two distinctive subgroups.

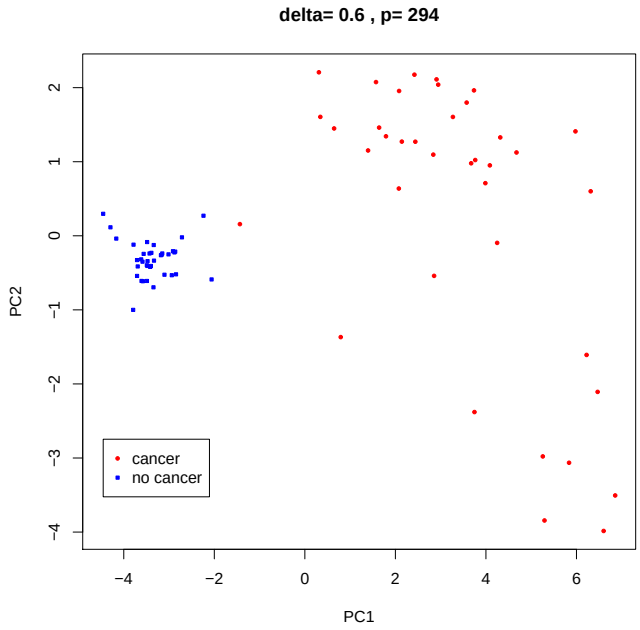

Figure A.35: TCGA: Using 294 methylation sites, i.e. with $\delta \geq 0.60$, it shows two distinctive subgroups. The cancerous cluster is more spreadout.

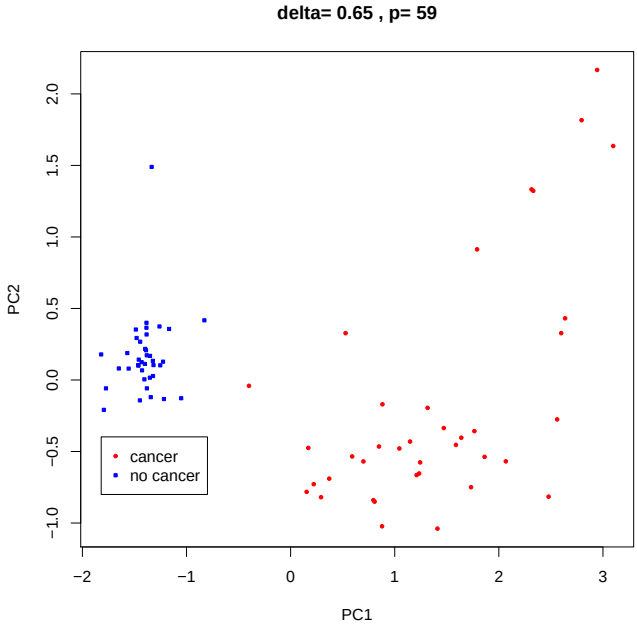

Figure A.36: TCGA: Using 59 methylation sites, i.e. with $\delta \geq 0.65$, the two clusters are moving closer together.

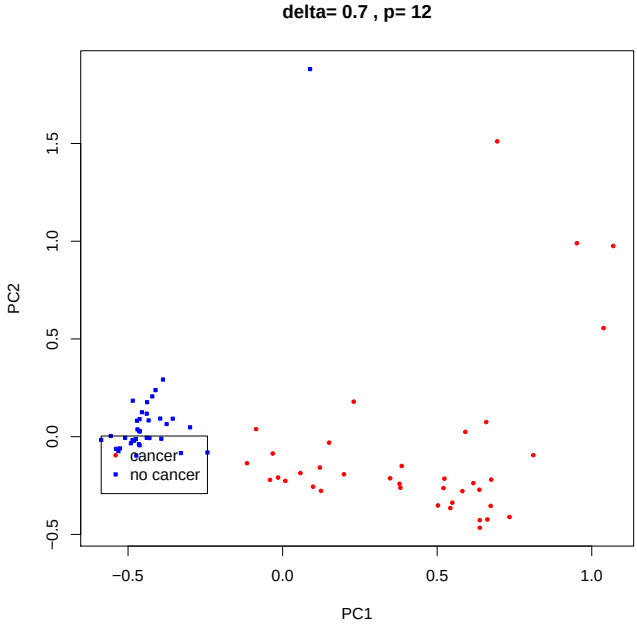

Figure A.37: TCGA: Using 12 methylation sites, i.e. with $\delta \geq 0.70$, it is difficult to separate the two clusters.

Table A.4: TCGA: The columns represent the δs/number of methylation sites for the MDA and the rows represent the δs/number of methylation sites for the difference in mean of the cancerous and non-cancerous subgroups. The result in each cell stands for the number of methylation sites captured/proportion captured by MDA.

	0.7/ 12	0.6/ 294	0.5/ 2915	0.4/ 16997	0.3/ 59954	0.2/ 135557	0.1/ 223327	0.0/ 385885
0.7/ 2	1/ 50%	2/ 100%	2/ 100%	2/ 100%	2/ 100%	2/ 100%	2/ 100%	2/ 100%
0.6/ 43	1/ 2.3%	43/ 100%	43/ 100%	43/ 100%	43/ 100%	43/ 100%	43/ 100%	43/ 100%
0.5/ 495	2/ 0.4%	104/ 21%	477/ 94.4%	494/ 99.8%	495/ 100%	495/ 100%	495/ 100%	495/ 100%
0.4/ 3488	2/ <0.1%	128/ 3.7%	1143/ 32.8%	3392/ 97.2%	3488/ 100%	3488/ 100%	3488/ 100%	3488/ 100%
0.3/ 14347	2/ <0.1%	148/ 1.0%	1488/ 10.4%	6982/ 48.7%	14128/ 98.5%	14347/ 100%	14347/ 100%	14347/ 100%
0.2 40878	2/ <0.1%	168/ <0.1%	1770/ <0.1%	9240/ 22.6%	27445/ 67.1%	40555/ 99.2%	40878/ 100%	40878/ 100%
0.1/ 94299	3/ <0.1%	203/ <0.1%	2173/ <0.1%	19914/ 12.6%	39365/ 41.7%	78118/ 82.8%	93754/ 99.4%	94299/ 100%
0.0/ 385885	12/ <0.1%	294/ <0.1%	2915/ <0.1%	16997/ 4.4%	59954/ 15.5%	135557/ 35.1%	223327/ 57.9%	385885/ 100%

was related to the shift in the mean of the methylation β-values for each sites, we could estimate the relevant methylation sites captured by the Gaussian mixture model scheme by calculating the proportion of the methylation sites that had the biggest difference in mean β-values captured by MDA.

As shown in Table A.4, there were overlaps in the table indicating that the Gaussian mixture modeling can pick up relevant CpG sites. In fact, if the difference of mean β-value population subgroups was 0.15, MDA actually picked up a majority of these sites. Although Gaussian mixture modeling did not identify the subgroups, it picked up enough relevant methylation sites so that clustering was possible unless the difference is small (< 0.05). The algorithm was tolerant of noises - up to an additional 6000 non-related

CpG sites before the signal disappeared.

A.3 STEPP Subpopulation Analysis for Continuous, Binary and Count Outcomes

A.3.1 Additional Aspirin Analysis Results: placebo vs 325 mg

The result of the original study was published in the New England Journal of Medicine (Baron et al., 2003) and concluded that low-dose aspirin has a moderate chemopreventive effect on adenomas (including lesion) in the large bowel. We used STEPP to investigate whether the magnitude of the treatment effect is similar across subpopulations defined by patient age. The first STEPP analysis comparing placebo vs 81 mg of daily aspirin was presented in Section 3.4.2. Here is the second analysis comparing placebo vs 325 mg of daily aspirin.

The GLM model within each subpopulation can be written as

$$logit(p) = \beta_0 + \beta_1 \times trt \tag{3.1}$$

The treatment indicator, trt, is 1 if it is 325 mg daily aspirin and 0 if it is placebo. The covariate of interest is age, which is treated as a continuous variable. The STEPP subpopulations are created by setting $r2$ to be 100 and $r1$ to be 30. Based on this setting, 8 subpopulations are generated. The subpopulations summary can be found in Table 3.3.

Based on these 8 subpopulations, treatment effect estimates of subpopulations are computed and the resulting STEPP plots are generated (see Figure A.38, Figure A.39 and Figure A.40). The result is not significant indicating that there is unlikely an interaction between treatment and age groups.

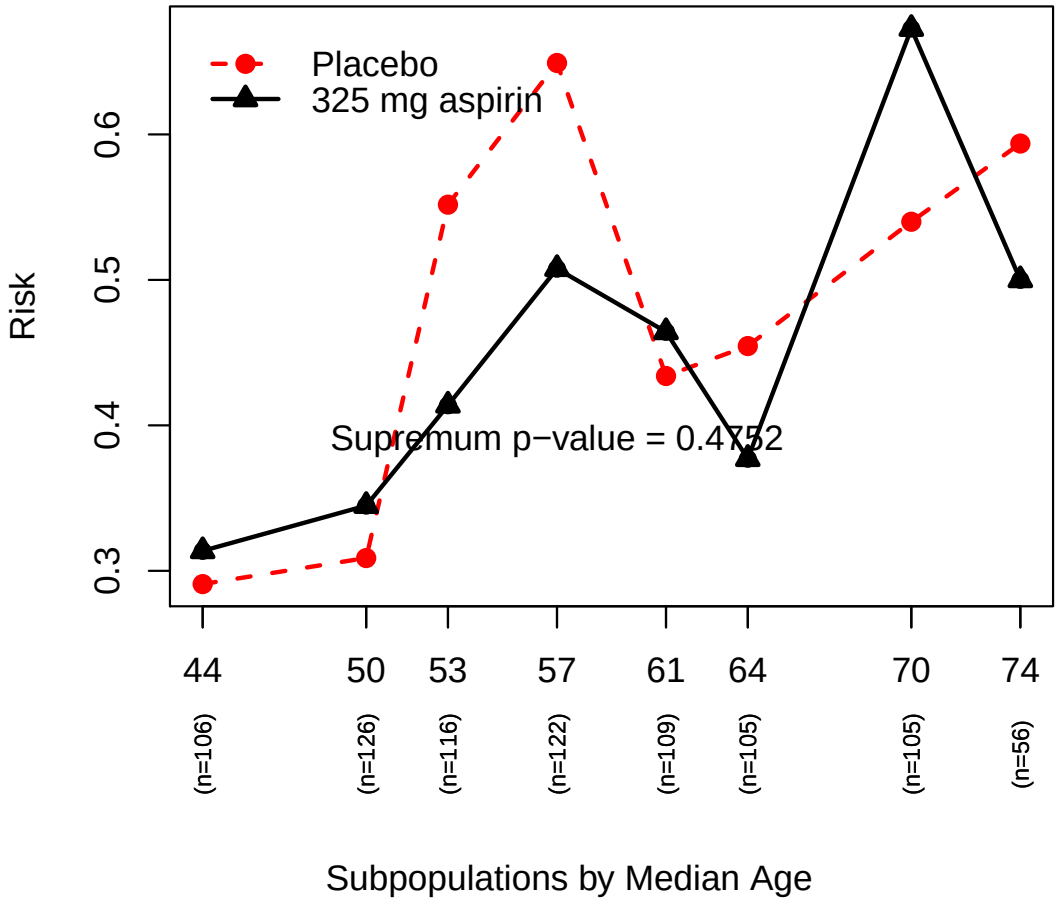

Figure A.38: The STEPP plot shows the absolute risk (or probability of experiencing AD) for two treatment groups across different age subgroups - the ″red″ dashed line is the placebo group and the ″black″ solid line is the 325 mg aspirin group.

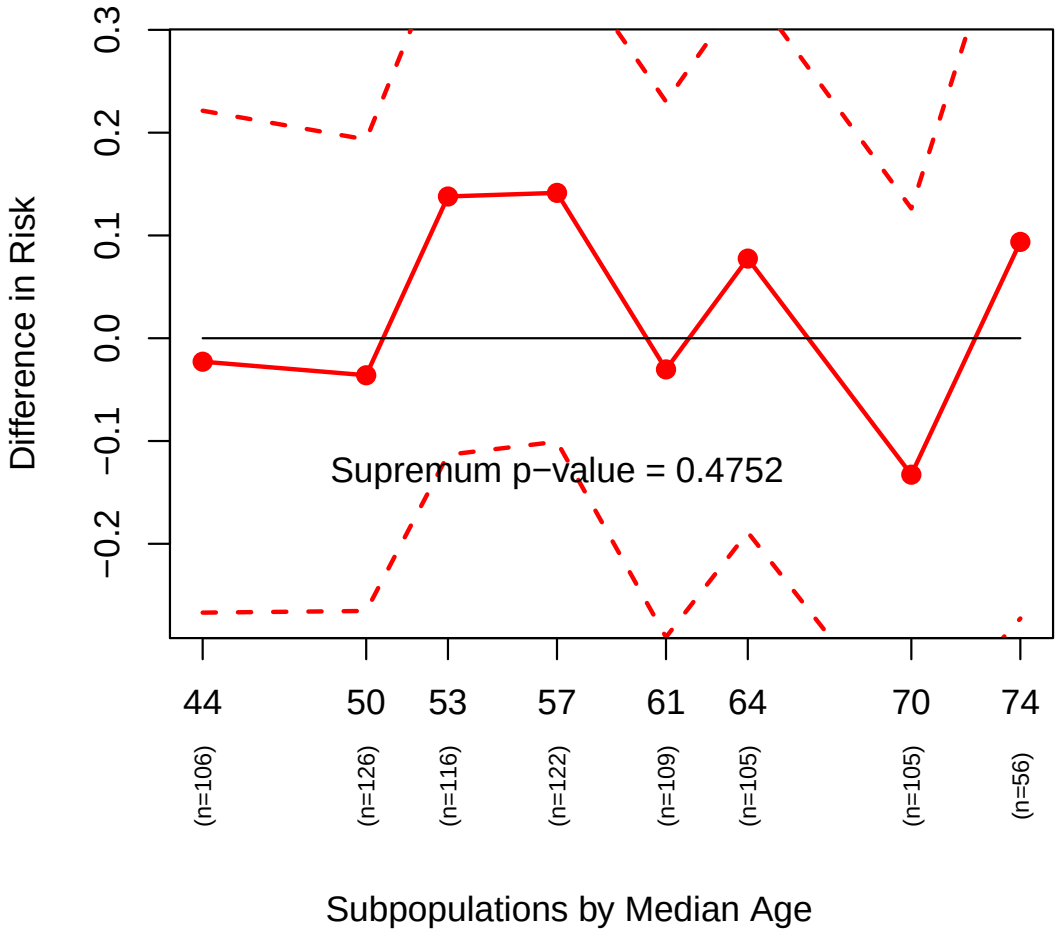

Figure A.39: The STEPP plot shows the differences in risk of experiencing AD across the various age subgroups between placebo and the 325 mg aspirin treatment groups. The interaction supremum p-value based on risk difference is 0.48, suggesting an insignificant result.

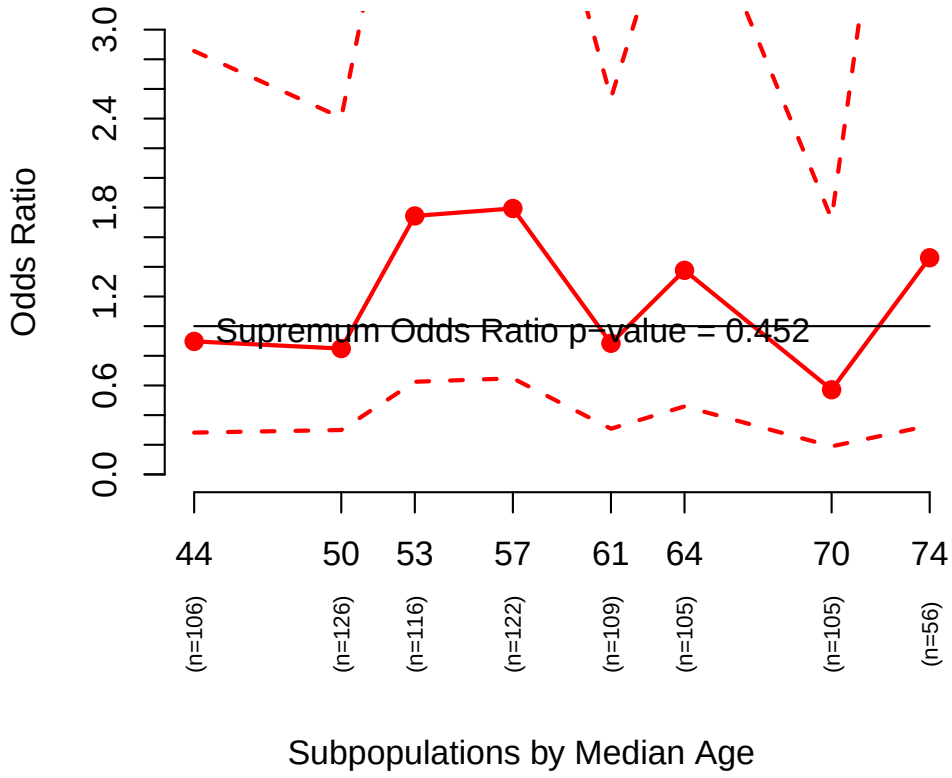

Figure A.40: The STEPP plot shows the odds ratio of experiencing AD across the age subgroups between placebo and the 325 mg aspirin treatment groups. The overall odds ratio of experiencing AD is about 1.1 when comparing the two groups. The interaction supremum p-value based on odds ratio is 0.452, also suggesting an insignificant result.

A.3.2 Additional Aspirin Analysis Results: 81 mg vs 325 mg

Here is the third analysis comparing 81 mg vs 325 mg of daily aspirin.

The GLM model within each subpopulation can be written as

$$logit(p) = \beta_0 + \beta_1 \times trt \tag{3.2}$$

The treatment indicator, trt, is 1 if it is 325 mg daily aspirin and 0 if it is 81 mg. The covariate of interest is age, which is treated as a continuous variable. The STEPP subpopulations are created by setting $r2$ to be 100 and $r1$ to be 30. Based on this setting, 8 subpopulations are generated. The subpopulations summary can be found in Table 3.3.

Based on these 8 subpopulations, treatment effect estimates of subpopulations are computed and the resulting STEPP plots are generated (see Figure A.41, Figure A.42 and Figure A.43). The result is borderline significant indicating that there is possible interaction between treatment and age groups.

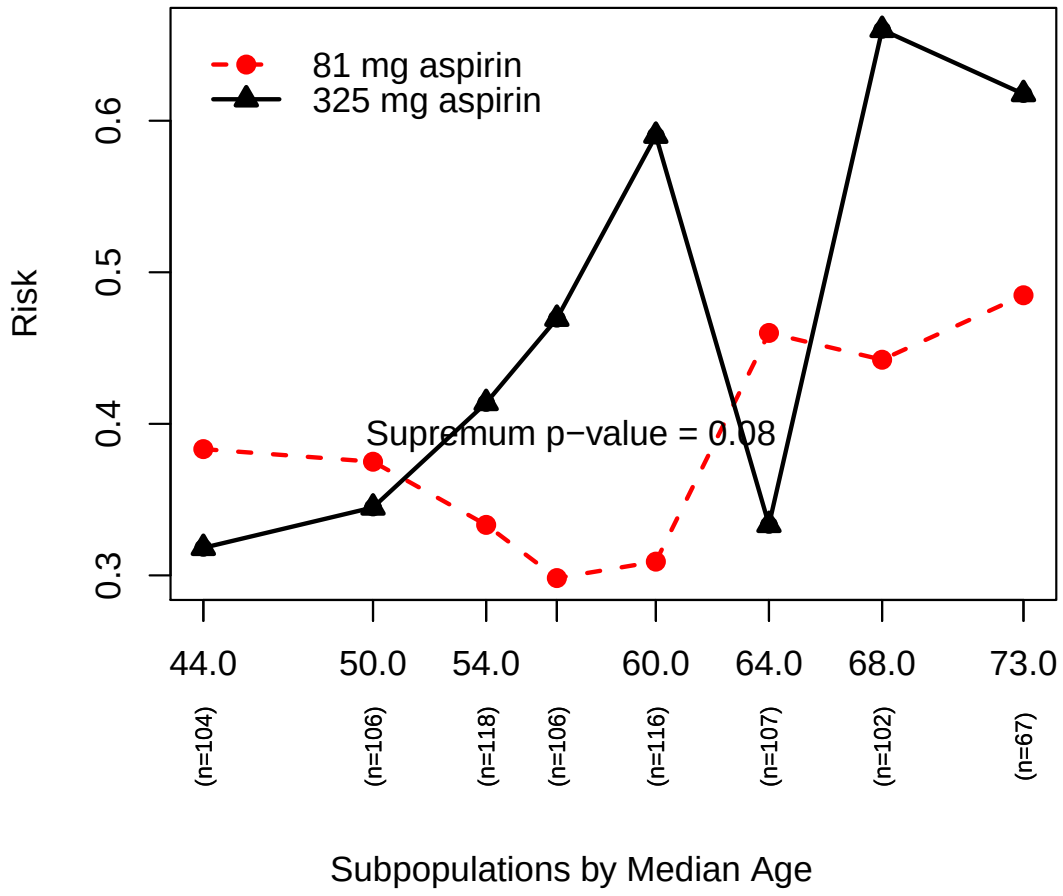

Figure A.41: The STEPP plot shows the absolute risk (or probability of experiencing AD) for two treatment groups across different age subgroups - the ″red″ dashed line is the 81 mg aspirin group and the ″black″ solid line is the 325 mg aspirin group.

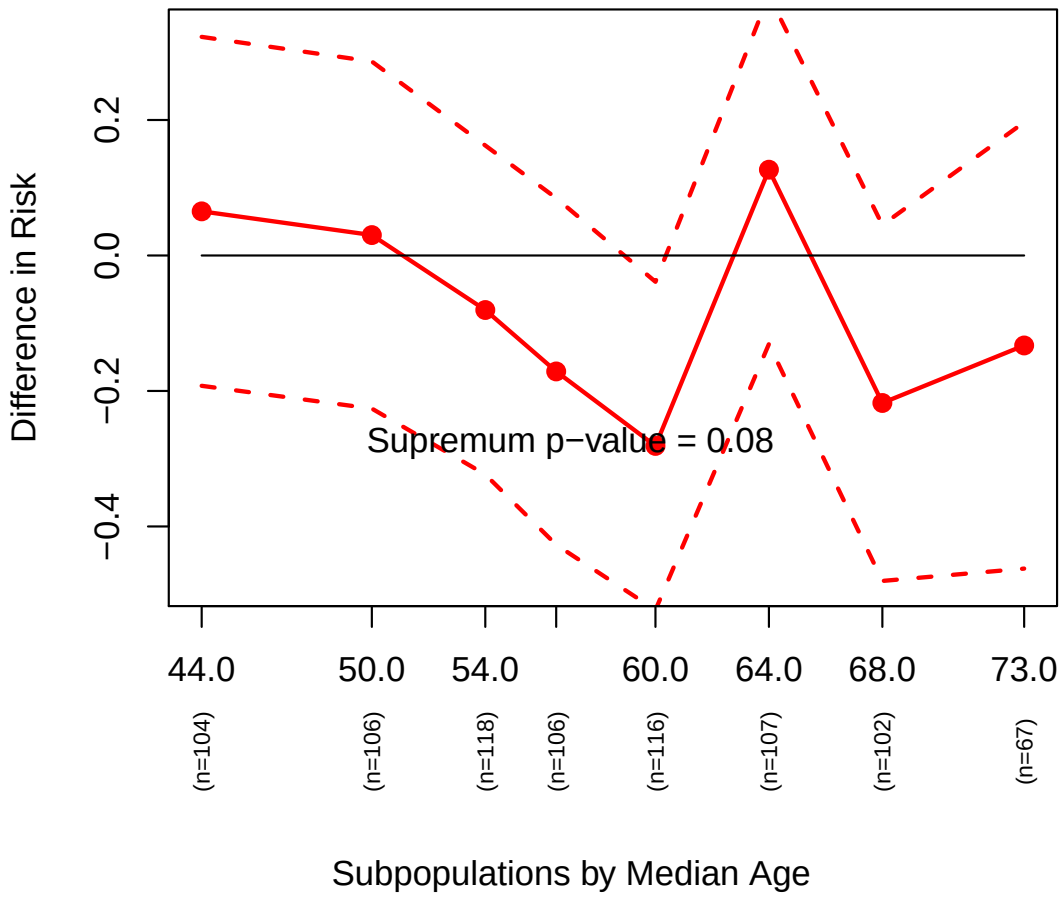

Figure A.42: The STEPP plot shows the differences in risk of experiencing AD across the various age subgroups between 81 mg and the 325 mg aspirin treatment groups. The interaction supremum p-value based on risk difference is 0.08, suggesting a borderline significant result.

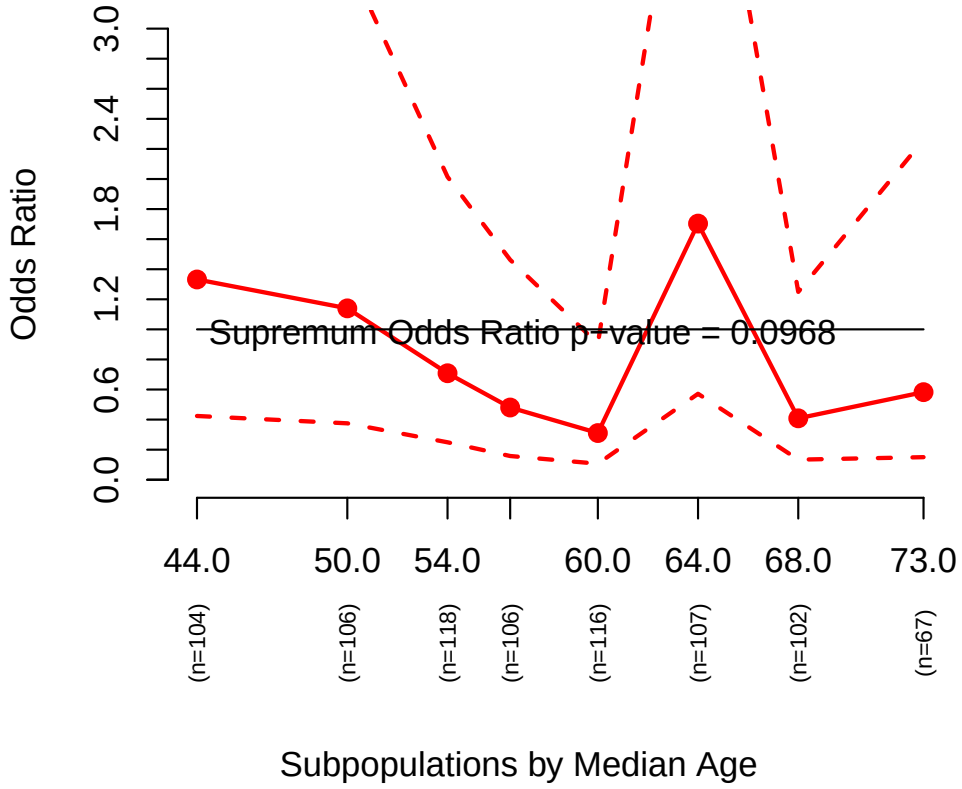

Figure A.43: The STEPP plot shows the relative risk of experiencing AD across the age subgroups between 81 mg and the 325 mg aspirin treatment group. The overall odds ratio of experiencing AD is about 0.75 when comparing the two groups. The interaction supremum p-value based on odds ratio is 0.097, also suggesting a possible borderline interaction effect between risks and the age-defined subpopulations on the relative scale.

A.3.3 Additional Simulation Details

The following are the results of all the simulations done for the different Gaussian, Binomial and Poisson models under the null. As reported in Section 3.4, some of the permutation tests produce slightly inflated Type I error probability (α). But, as the sample size

n increases, as one would expect, the results shrink towards the nominal Type 1 error rate for all these models.

Table A.5: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Gaussian model N(55,49). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.004	0.052	0.118
			T2	0.018	0.066	0.120
			T3	0.012	0.038	0.090
			T1*	0.004	0.052	0.116
			T3*	0.016	0.044	0.090
200	60	80	T1	0.018	0.050	0.092
			T2	0.010	0.064	0.118
			T3	0.012	0.054	0.120
			T1*	0.016	0.054	0.096
			T3*	0.012	0.058	0.112
500	150	200	T1	0.006	0.046	0.086
			T2	0.014	0.038	0.096
			T3	0.012	0.056	0.088
			T1*	0.006	0.044	0.088
			T3*	0.012	0.054	0.092
1000	300	400	T1	0.010	0.054	0.106
			T2	0.010	0.048	0.108
			T3	0.010	0.062	0.108
			T1*	0.010	0.056	0.104
			T3*	0.010	0.062	0.106

Table A.6: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Gaussian model N(75,25). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.014	0.066	0.104
			T2	0.012	0.058	0.116
			T3	0.014	0.054	0.112
			T1*	0.014	0.066	0.106
			T3*	0.014	0.058	0.110
200	60	80	T1	0.010	0.046	0.122
			T2	0.008	0.052	0.104
			T3	0.004	0.042	0.104
			T1*	0.010	0.050	0.124
			T3*	0.008	0.044	0.096
500	150	200	T1	0.014	0.046	0.096
			T2	0.012	0.044	0.084
			T3	0.010	0.046	0.108
			T1*	0.014	0.048	0.092
			T3*	0.010	0.048	0.124
1000	300	400	T1	0.006	0.046	0.096
			T2	0.014	0.058	0.120
			T3	0.016	0.056	0.120
			T1*	0.006	0.046	0.096
			T3*	0.014	0.058	0.124

Table A.7: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Binomial model Bin(n,0.3). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.014	0.056	0.102
			T2	0.014	0.058	0.098
			T3	0.010	0.048	0.098
			T1*	0.012	0.054	0.098
			T3*	0.016	0.050	0.120
200	60	80	T1	0.016	0.040	0.102
			T2	0.012	0.066	0.114
			T3	0.008	0.046	0.096
			T1*	0.006	0.050	0.108
			T3*	0.004	0.044	0.098
500	150	200	T1	0.002	0.044	0.098
			T2	0.016	0.062	0.106
			T3	0.006	0.062	0.100
			T1*	0.006	0.050	0.096
			T3*	0.006	0.066	0.120
1000	300	400	T1	0.012	0.044	0.102
			T2	0.010	0.054	0.102
			T3	0.014	0.044	0.090
			T1*	0.010	0.052	0.102
			T3*	0.006	0.050	0.100

Table A.8: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Binomial model Bin(n,0.5). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.012	0.058	0.100
			T2	0.010	0.054	0.098
			T3	0.012	0.062	0.106
			T1*	0.012	0.046	0.098
			T3*	0.010	0.062	0.110
200	60	80	T1	0.020	0.052	0.098
			T2	0.014	0.056	0.094
			T3	0.022	0.046	0.092
			T1*	0.020	0.060	0.100
			T3*	0.014	0.060	0.104
500	150	200	T1	0.010	0.048	0.114
			T2	0.004	0.040	0.092
			T3	0.004	0.044	0.098
			T1*	0.010	0.048	0.112
			T3*	0.008	0.042	0.094
1000	300	400	T1	0.010	0.056	0.124
			T2	0.016	0.052	0.112
			T3	0.020	0.056	0.100
			T1*	0.010	0.054	0.120
			T3*	0.018	0.056	0.122

Table A.9: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Binomial model Bin(n,0.7). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.008	0.046	0.104
			T2	0.016	0.060	0.128
			T3	0.018	0.056	0.108
			T1*	0.014	0.046	0.102
			T3*	0.018	0.048	0.112
200	60	80	T1	0.010	0.050	0.090
			T2	0.010	0.054	0.114
			T3	0.012	0.056	0.110
			T1*	0.006	0.058	0.100
			T3*	0.012	0.066	0.118
500	150	200	T1	0.006	0.042	0.096
			T2	0.010	0.046	0.102
			T3	0.008	0.048	0.082
			T1*	0.010	0.042	0.094
			T3*	0.006	0.048	0.110
1000	300	400	T1	0.006	0.054	0.100
			T2	0.010	0.050	0.098
			T3	0.010	0.040	0.114
			T1*	0.014	0.060	0.104
			T3*	0.008	0.048	0.100

Table A.10: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Poisson model Pois(5). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.006	0.044	0.096
			T2	0.006	0.040	0.084
			T3	0.004	0.048	0.092
			T1*	0.008	0.040	0.102
			T3*	0.008	0.044	0.084
200	60	80	T1	0.004	0.042	0.104
			T2	0.008	0.046	0.100
			T3	0.008	0.050	0.106
			T1*	0.006	0.044	0.096
			T3*	0.012	0.056	0.098
500	150	200	T1	0.008	0.042	0.082
			T2	0.006	0.052	0.114
			T3	0.010	0.046	0.104
			T1*	0.012	0.040	0.080
			T3*	0.008	0.040	0.086
1000	300	400	T1	0.018	0.046	0.100
			T2	0.008	0.062	0.108
			T3	0.006	0.040	0.094
			T1*	0.020	0.046	0.100
			T3*	0.004	0.066	0.102

Table A.11: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Poisson model Pois(10). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.010	0.048	0.086
			T2	0.008	0.046	0.094
			T3	0.004	0.048	0.074
			T1*	0.012	0.050	0.080
			T3*	0.002	0.044	0.098
200	60	80	T1	0.012	0.052	0.104
			T2	0.010	0.052	0.112
			T3	0.010	0.048	0.120
			T1*	0.014	0.050	0.110
			T3*	0.012	0.052	0.100
500	150	200	T1	0.008	0.050	0.108
			T2	0.010	0.064	0.124
			T3	0.014	0.062	0.110
			T1*	0.008	0.054	0.114
			T3*	0.010	0.060	0.116
1000	300	400	T1	0.014	0.042	0.088
			T2	0.008	0.054	0.104
			T3	0.002	0.040	0.102
			T1*	0.014	0.040	0.096
			T3*	0.004	0.054	0.102

Table A.12: Estimated α level of the permutation test for interaction based on the statistics T1, T2, T3, T1*, and T3* as defined in Section 3.3.2 with outcome Y under the Poisson model Pois(15). The distribution of the covariate of interest, Z, is N(25,100). Results are based on 500 simulations of sample size n, with subpopulation generating parameters r1 and r2.

				α		
n	r1	r2	statistic	0.01	0.05	0.10
100	30	40	T1	0.010	0.050	0.108
			T2	0.016	0.056	0.106
			T3	0.006	0.042	0.098
			T1*	0.008	0.062	0.106
			T3*	0.008	0.052	0.096
200	60	80	T1	0.006	0.060	0.108
			T2	0.006	0.050	0.100
			T3	0.012	0.042	0.102
			T1*	0.006	0.056	0.110
			T3*	0.012	0.046	0.106
500	150	200	T1	0.008	0.042	0.108
			T2	0.008	0.052	0.104
			T3	0.008	0.058	0.108
			T1*	0.008	0.046	0.106
			T3*	0.004	0.040	0.094
1000	300	400	T1	0.006	0.046	0.110
			T2	0.016	0.056	0.104
			T3	0.012	0.054	0.106
			T1*	0.006	0.046	0.110
			T3*	0.004	0.052	0.106