GENETIC ASSOCIATION ANALYSIS ON SECONDARY PHENOTYPES AND GROUP CONDITIONAL VARIABLE IMPORTANCE IN OPPERA STUDY

Wei Xue

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

Eric Bair

Mengjie Chen

Yun Li

Smith Shad

Donglin Zeng

ABSTRACT

Wei Xue: Genetic Association Analysis on Secondary Phenotypes and Group Conditional Variable Importance in OPPERA Study
(Under the direction of Eric Bair)

Temporomandibular disorder (TMD) is a complex chronic painful orofacial disorder resulting in dysfunction in the temporomandibular joints and the muscles around the jaw. Numerous risk factors were studied and identified for the chronic and onset of TMD. The Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study is a prospective study designed to study the etiology and the risk factors contributing to the onset and chronic TMD (Smith et al. 2011).

Genetic risk factors play an important role in the etiology of TMD. While many studies identified the genetic variants associated with TMD case-control status, one may wish to identify genetic markers associated with secondary phenotypes (such as clinical pain) that are related to the severity of TMD. In such cases, naive regression methods that ignore the case-control design produces biased results. This problem may be corrected by statistical methods such as inverse probability weighting (IPW). However, it may be unreliable when genetic markers and secondary phenotypes are strongly associated with case-control status. In order to perform unbiased association analysis, we proposed a novel permutation-based IPW method, and compared it with conventional IPW method. The results indicated that the permutation-based IPW produced controlled type I error rates with no loss in power. The application to the data from OPPERA study identified the associated SNPs with the severity of orofacial pain.

Numerous risk factors were studied in previous OPPERA studies to cast light on the etiology of TMD. It is of great interest to researchers to know if a subset of variables have

high variable important (VIMP) score conditional on existing risk factors. The are curious to know that in addition to the existing variables, would the group variables bring more information in predicting outcome. For example, they want to know if the group importance score for measuring mechanical and thermal pain sensitivity is significantly different from 0 conditional on all the other variables when predicting either chronic or first-onset TMD.

In the second project, we proposed a method to test the group conditional variable importance statistically, by conditional distribution of group variables on the those outside the group using random forest model. Simulations were performed by continuous and categorical variable types. p-values were calculated for some groups. This method corrects the shortcomings of the likelihood of choosing correlated variables with spurious correlation, and provided a way of testing group variables without bias.

The methodology described in the second topic was applied to data in OPPERA case-control and cohort study in the third topic for both chronic and first-onset TMD. Correlated risk factors were subset and tested by the proposed method for the null hypothesis of group VIMP score not significantly different from 0 based on the rest risk factors of TMD in the data set. A number of groups of variables were identified bringing more information for the chronic TMD in addition to the existing risk factors. But none of the proposed groups were identified conditionally important in first-onset TMD study.

To Yanping Shi and Zhixin Xue, my parents who give me unconditional love and support.
To Kai Xia, my husband and my best friend.
To Yo-Yo and Mo-Mo, my sweet kids .

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my DrPH advisor Dr. Eric Bair for his fully unwavering support and mentorship in my dissertation. This dissertation cannot be finished without his patience, wisdom, understanding and guidance. I would also like to express my appreciation to my committee members Dr. Donglin Zeng, Dr. Yun Li, Dr. Mengjie Chen and Dr. Shad Smith for their valued comments.

I would also like to thank my parents and my in-laws. Without their unconditional love and help to taking care of my kids and me, I cannot finish the dissertation on time.

I would like to dedicate my dissertation to my beloved kids Yo-Yo and Mo-Mo. They are the angel in my life. Their little smile shed lights on me, and give me the power to move on.

Finally, I would like to give my special thank my husband and best friend Kai Xia. He is an amazing husband and great father. Without his love and willingness to support me unconditionally, taking care of kids and doing the housework, the dissertation would have taken longer to finish. We took our hands together in the past 10 years and our life together will be more amazing in the rest of the journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Chronic pain is a significant and costly health problem . In particular, temporomandibular disorder (TMD) is a complex chronic painful orofacial disorder resulting in dysfunction in the temporomandibular joints and the muscles around the jaw. It has a prevalence of around 5% in adults in US 2012 National Health Interview survey (Isong et al. 2008). Numerous risk factors were studied and identified for the chronic and onset of TMD, such as pain sensitivity risk factors, clinical risk factors, psychological distress, and genetic factors. The Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study is a prospective study designed to study the etiology and find out the risk factors contributing to the onset and chronic TMD (Smith et al. 2011). The study patients filled out questionnaires and underwent a series of clinic examinations for potential risk factors for TMD. Blood sample was drawn from each participants and genotyped for genome-wide association study (GWAS).

There are compelling evidence that genetic risk factors may play a role in the etiology of TMD. A recent study by Plesh et al. (2012) illustrated that genetic risk factors partially contributed to TMD pain in women. While many studies identified the genetic variants associated with TMD case-control status and they are informative, one may wish to identify genetic markers associated with secondary phenotypes (such as clinical pain) that are related to the severity of TMD. In such cases, naive regression methods that ignore the case-control design will produce biased results. This problem may be corrected by statistical methods such as inverse probability weighting (IPW), which assigns weights to subjects based on case-control status to correct for the fact that cases are overrepresented in

a case-control study (Richardson et al. 2007) and (Monsees et al. 2009). However, conventional IPW regression may be unreliable when evaluating the association between genetic markers and secondary phenotypes that are strongly associated with case status. In a TMD case-control study, one may wish to identify genetic markers associated with the severity of orofacial pain, which is present in all TMD cases, but nearly all controls report no orofacial pain, causing conventional IPW regression to produce inflated type I errors and inaccurate results. In order to perform the association analysis, we proposed a novel permutation-based IPW method in the first project, and compared it with conventional IPW method. The results from simulations indicated that whereas conventional IPW method might produced inflated type I error rates, the permutation-based IPW produced correct type I error rates with no loss in power. We then applied this method to data from OPPERA case-control study to identify the associations between candidate SNPs and the severity of orofacial pain. Two novel SNPs were identified associated with TMD pain severity by our method.

Numerous risk factors were studied in previous OPPERA researches to cast light on the etiology of TMD. Univariate models or ANOVA model were utilized to find out the putative risk factors in chronic and first-onset TMD. Despite the informative studies, one may wish to identify the most vital risk factors in predicting TMD. Random forest modeling was utilized to identify the important risk factors by Bair et al. (2013b). But statistical tests for the variable importance cannot be obtained by the previous random forest model in OPPERA study. Breiman (2001) and Strobl et al. (2007) provided methods that tend to assign high variable important scores to the correlated variables which indeed not associated with the outcome. In order to solve the issue, our group previouly developed a method based on the conditional distribution of variable of interest, and statistically tested the null hypothesis that the conditional VIMP score for a risk factor is 0. This method

successfully avoid choosing spurious variables as important ones. Based on the methodology, people may be curious to know whether a group of variables are significantly important to the outcome (which might be chronic or first-onset TMD) conditional on the rest risk factors. Researchers in OPPERA study are especially interested in such questions because of the large amount of variables and the high correlations among them in OPPERA case-control study and cohort study. For example, one may wish to know if the group importance score for measuring mechanical and thermal pain sensitivity is significantly different from 0 when predicting either chronic or first-onset TMD conditional on the other existing variables.

In the second project, we proposed a method to statistically test the group conditional variable importance, by using conditional distribution of variables in the group on the those outside the group in random forest. The idea behind the scene is when the variables in the group are important conditional on the other variables, they would bring more information significantly in predicting the outcome in addition to variables outside the group. The the prediction error by original data using random forest would be significantly different from that by data of replacing the group with conditional distribution of group variables on the rest of variables. Simulations were performed by different variable types. p-values were calculated for some risk factors groups in both chronic and first-onset TMD data sets. This method corrects the shortcomings of the likelihood of choosing correlated variables with spurious correlation, and provided a way of testing group variables without bias.

The methodology described in the second topic were applied to OPPERA case-control study and cohort study in the third topic for both chronic and first-onset TMD. The putative risk factors for TMD that were measured in OPPERA study were described in this section. Correlated risk factors were subset and tested by the proposed method for the null hypothesis that group VIMP score is not significantly different from 0. Some groups

of variables were identified bringing more information for the chronic TMD in addition to the other risk factors. But none of those groups were identified significantly important in first-onset TMD.

CHAPTER 2: LITERATURE REVIEW

2.1 Statistical Methods and Background Information in Secondary Phenotypes and Genetic Association Studies

2.1.1 Genetic Association Studies

Genetic association studies identifies associations between disease or phynotype and a set of genetic markers. The aim is to find the candidate genes or specific regions which might contribute to that trait Lewis and Knight (2012). The genetic association study is more powerful for detecting common complex disease than other methods, such as linkage analysis, which is based on the family data Austin et al. (2013). The diseases and traits are complex since they include both genetic factors and environmental factors. Also, the relationships between some traits and genes can only be tested through genetic association studies because they incorporate the traditional epidemiological design. Association studies may be performed on candidate genes or the full genome, which is known as a genome-wide association study (GWAS). They are used to test the null hypothesis of no association between a marker and a phenotype under the "common disease/common variant" (CDCV) hypothesis (Austin et al. 2013).

Candidate gene studies seek to identify variation within a particular gene or sets of genes that may be associated with a trait or disease of interest. It tries to determine if the allele in candidate genes are more frequently seen in cases (which might be disease under investigation) rather than controls (which do not have disease). Researchers need to understand the disease mechanisms to choose the possible candidate genes (Kwon and Goate

2000). Usually the first step is to look at the genes that are related to the disease or the trait in published studies and decide if they include variants with function (Tabor et al. 2002). For example, to select the candidate genes for alcoholism, we need to identify different genetic pathways related to the alcoholism mechanism and the related enzymes and chemicals based on human studies, animal models and the expression of genes in cells or tissues (Kwon and Goate 2000). After the candidate genes are chosen, researchers need to decide which polymorphism in the candidate gene to choose. The polymorphism represents the variation of the gene at one site in the population. These variation sites are called Single Nucleotide Polymorphism (SNPs). SNPs are the polymorphisms used most often in the genetic association study because they occur frequently in the genome, and are relatively easier to genotype. In addition to SNPs, other structural variants such as insertions and deletions, translocations, VNTRs (variable number of tandem repeats) and some other type of polymorphisms can be used in genetic studies (Austin et al. 2013). When the possible genes and polymorphisms are chosen, researchers will test them in a case-control study where random samples with or without diseases are included. The advantage of the candidate genes approach is that it will quickly evaluate the possible associations between traits and genes with less concern about spurious correlations due to multiple comparisons (Kwon and Goate 2000).

Genome-wide association studies (GWAS) are one tool for evaluating the association between phenotype and related genes. The objective is to identify the genetic variants associated with a specifec phenotype by investigating SNPs across the whole genome. The SNP and CNV calling can be performed using genotyping arrays or next-generation sequencing technology followed by quality control analysis from blood or buccal samples in the study. There is no prior hypothesis that SNPs are correlated with disease. A much larger number of SNPs are involved in GWAS because millions of SNPs are read using high-throughput technology.

GWAS studies may use cohort, case-control, and case-parent trio designs. In case-control studies, people with disease (cases) and without disease (controls) are compared with respect to millions of SNPs. When a SNP is associated with specific trait, the allele frequency is significantly different in the cases compared with controls. Large samples are needed in order to have sufficient statistical power to detect such associations. Researchers should also be aware of multiple testing issues that might cause false positives. The strength of the association between the SNPs and the specific trait are usually estimated as odds ratios (OR) based on logistic regression models. QQ-plots and Manhattan plots can be used to display the significant findings (Austin et al. 2013). Linear regression may be performed when the outcome is continuous (Cantor et al. 2010).

2.1.2 The analsyis of secondary phenotype in genetic association studies

Prospective studies are the preferred study design for evaluating the relationship between exposure and outcome. However, prospective studies are time-consuming and expensive, especially for rare disease, because more participants needs to be enrolled in the study to have adequate statistical power. Since genotyping can be expensive, prospective GWAS studies are generally not feasible. Case-control studies recruit a number of subjects with and without a phenotype of interest. Most GWAS studies use case-control design since it is faster and cheaper than a prospective study. And researchers are interested in the relationship between a given SNP and disease case-control status.

Given the expense required to perform a GWAS study, it is common to collect more information from each subject in addition to the disease status, such as secondary phenotypes, to maximize the return. Secondary phenotypes are information in the subject related to the disease of interest (Ghosh et al. 2013), which helps to understand disease etiology and provide more information. For example, body mass index (BMI) and physical

activity are of great interest to researchers in terms of the association of secondary phenotype and genotype (Frayling et al. 2007), which helped them to understand the association of diabetes and genetic variants as well. Other examples include height in hypertensions study (Loos et al. 2008), lipid (high-density lipoproten cholesterol, low-density lipoproten cholesterol) in coronary artery disease study (Kammerer et al. 2004), and ages of menarche in breast cancer GWA study (He et al. 2009)

To study the association of secondary phenotype and genetic variants, it is common utilized methods such as standard linear regression for quantitative outcome, or logistic regression for categorical secondary phenotype. However, the association studies between secondary traits and genetic variants are complicated because of the case-control sampling scheme. Case-control studies are not a representative sample from the population, meaning that estimates derived from case-control samples may be biased (de Dieu Tapsoba et al. 2014). Standard linear regression analysis is typically used to evaluate the association between a quantitative phenotypes and a given genetic marker. When evaluating the association between a secondary quantitative phenotype and a marker, it is common to analyze only cases or controls, or combine both cases and controls but ignore their case-control status, or use meta-analysis for cases and controls, or analyze both cases and controls by adjusting for the case-control status (Lin and Zeng 2009). However, these approaches may produce biased regression estimates. Several statistical methods have been proposed to correct for the case-control sampling scheme. Richardson et al. (2007) and Monsees et al. (2009) investigated that inverse-probability weighting (IPW) avoided this issue by taking selection probability into calculation. Lin and Zeng (2009) proposed method using maximum likelihood estimation to analyize secondary phenotype under different scenarios including rare disease, and under different disease rates. Although it is powerful, the assumptions of the distribution of secondary phenotype is needed. He et al. (2011) performed the analysis of genetic association with secondary phenotype using Gaussian

copula method where multiple correlated secondary phenotypes were able to handle. Rare diseases were also studied with respect to the secondary phenotype. Li et al. (2010) found that standard approach produced biased results in rare disease when both secondary phenotype and genetic variants interact with primary disease outcome.

IPW is frequently used for the analysis of the association between secondary traits and genetic variants (Ghosh et al. 2013). IPW assigns weights to each study participant to adjust for the case-control study design. The participants are weighted so that the weight of cases in the study is comparable to the proportion of cases in the general population. Consider the following example: Suppose there are n_1 cases and n_0 controls in the study and the population prevalence of the diseas is s. Suppose the weight of controls are 1. Then cases are given a weight of $n_0 \times s/(n_1 \times (1-s))$, so that the weights of the cases in the study are the same as if one sampled from the general population. Appropriate methods are needed to estimate the standard error of regression coefficients after applying IPW (Monsees et al. 2009). The power for the IPW method is not as high as the power of unadjusted analysis, but the unadjusted methods has inflated type I error, especially when there are associations between genotype and secondary traits or between disease and genetic covariates (Monsees et al. 2009). Compared with other methods such as maximum likelihood estimation, IPW regression is more robust and flexible in terms of model misspecification (de Dieu Tapsoba et al. 2014).

2.1.3 Principal component analysis (PCA)

Principal component analysis (PCA) is a useful technique for multivariate data reduction that performs orthogonal transformations on a set of variables to create a new set of independent variables (Smith 2002). These new variables are called principal components. Usually the number of the principal components are less than or equal to the number of original variables. The first principal component has the largest variance among all the

components, and the succeeding components explains higher variance than those that follow them.

In genetic association studies, population stratification can produce spurious associations when the samples are from more than one population (Tian et al. 2008). In a case-control GWAS study, it is assumed that cases and controls are from the same population. However, when they come from different populations, the assumption is violated, which can produce false positive results (Liu et al. 2013). Population stratification results in systematic allele frequency differences in cases and controls. Thus, association between a marker and the outcome of interest is confounded by these systematic allele frequency differences between the two populations. Since population stratification can produce false positive results or reduce the power to detect real effects, there are methods to correct for population stratification, such as principal component analysis (PCA), linear mixed models (LMM), genomic control, and multidimensional scaling (Price et al. 2006, Li and Yu 2008, Zhang et al. 2010).

One of the most common methods for correcting bias from population stratification is principal component analysis (PCA), proposed by Price et al. (2006). His method, EIGENSTRAT, consists of 3 steps. PCA is performed on the data with the expectation that the largest principal components capture the systematic differences in allele frequencies due to population stratification. Linear regressions are performed between each SNP and first several components, and between outcome and the components for the residualized predictors and outcome. The (adjusted) association between the outcome and the genetic marker is evaluated by performing regression on the residuals of these regression models.

The advangtages of EIGENSTRAT include simplicity and computational tractability, especially for large GWAS data sets. It has greater power to detect true associations than the genomic control approach (Price et al. 2006). There are also some disadvantages to EIGENSTRAT. When there are discrete subpopulations in the data set, they may not

be adequately measured by continuous eigenvectors. This can produce spurious findings if there are outliers in the data (Liu et al. 2013). PCA also assumes the samples are independent. Thus, it is not appropriate for family-based studies or situations with cryptic relatedness, which can be analyzed using mixed effects models.

PCA were utilized to identify the putative latent constructs of risk factors in both chronic and first-onset TMD in Orofacial Pain Prospective Evaluation and Risk Assessment (OPPERA), such as psychosocial risk factors, pain sensitivity risk factors. It envolves four steps to do the PCA, including selecting variables for PCA, getting the correlation matrix, finding out principal component, and intepreting the factor loadings. In baseline case-control study, this method helped to identify that compare to chronic TMD cases, the controls are less sensitive to the stimulus in pain (Greenspan et al. 2011). This method also helped to identify four components which proved the correlation of potential psychosocial risk factors and chronic TMD (Fillingim et al. 2011). In first-onset TMD, PCA was performed to reduce the dimension and find out the putative latent construct in psychological risk factors. Four domains were accessed using this method, including Active Coping, Passive Coping, Global Psychological and Somatic Symptoms, as well as Stress and Negative Affectivity (Fillingim et al. 2013).

2.1.4 Permutation

A permutation is a random ordering of the elements of a set. Permutation testing provides a way to perform significance testing in GWAS studies by empirical test statistics. By randomly permuting the phenotype vector, the association between the phenotype and a given marker is removed, but the correlation among genotypes is preserved. Each permutation is a random sample from the original data. Since there is necessarily no association between the genetic markers and the permuted phenotype vector, the permutation procedure provides an estimate of the distribution of the test statistic under the null hypothesis

(Bush and Moore 2012). The p-value for testing the null hypothesis of no association between a genetic variant and the phenotype can be calculated by dividing the number of times when the permuted test statistic is more extreme than the test statistic on the original data by the number of permutations. Permutation is widely used jin the analysis of GWAS data and provides robust p-values when the assumptions of parametric models are violated (Posthuma et al. 2009).

2.2 Random Forests and Variable Importance

2.2.1 Random Forests Model

Random Forests are a machine learning technique widely used in many areas including genetics (Goldstein et al. 2010), ecology, (Cutler et al. 2007), physics, bioinformatics and may other fields. It is a non-parametric method, first introduced by Breiman in 2001, based on bagging, classification and regression trees (CART) (Breiman 2001). Bagging is a method for reducing the prediction variance by averaging a series of models (Hastie et al. 2009). Random forests have many desirable properties of decision trees and it is much more accurate.

The random forest algorithm is similar to bagging (bootstrap aggregation), which also involves growing a large number of decision trees. Every tree in the forest is fit to a bootstrap sample from the training data. In each node of the tree, only a subset of the predictors are considered when choosing the splitting variable in order to reduce the correlation between trees. The ensemble of trees is then used to produce estimates based on the forest. Predictions are obtained by averaging the predictions of each individual tree for regression problems and by majority vote for classification problems. The prediction error rate, which is similar to the cross validation error rate, is calculated based on the out of bag (OOB) samples in the random forest, which are the observations not included in the bootstrap sample for a given tree. The size of the tree is increased until the error rate is

stable. (Hastie et al. 2009).

Random forest has many good features. It can be used when the number of variables are large while the number of observations is small, known as "small n large p" problem. It can be applied to both categorical and continuous variables. Random forest is able to handle highly correlated predictors and account for arbitrary interactions. It also can model nonlinear associations between predictors and the outcome. In general they will not overfit the data. They can also be used to evaluate variable importance (Díaz-Uriarte and De Andres 2006).

Random forest models were performed in the previous OPPERA studies to identify the contributions of putative risk factors in the first-onset TMD analysis. It was utilized to identify the most important variables among the risk factors by assigning variable important score to the variable with the most important one with score 100. It is also performed to find out the correlation of the variable and first-onset TMD by partial dependence plot (Fillingim et al. 2013). Parafunctional oral behaviors, Pennebaker Inventory of Limbic Languidness (PILL) were identified as the most important varible in clinical risk factors and psychological risk factors respectively to the first-onset TMD (Fillingim et al. 2013, Ohrbach et al. 2013)

2.2.2 Variable Importance (VIMP) in Random Forest

Random forests have the capability of calculating the variable importance score of the predictors. Intuitively, a variable is "important" if the predictive accuracy of the model decreases if the variable is removed from the model or measured with error. There are multiple ways to measure variable importance. The naive method is to count the number of times each variable appears in all trees in the forest. A better approach is to permute a given variable and evaluate the decrease in the predictive accuracy of the model after the variable is permuted. We refer to this method as Breiman's variable importanace (VIMP).

The idea is that if a variable is not important, the predictive accuracy of the model will not change significantly after permuting it. A brief summary of the procedure is the following: to calculate the importance of X_j, variable X_j is permuated randomly in the OOB samples. The predictive accuracy of the model is calculated and compared to the accuracy of the model applied to the original data (with X_j unpermuted). The difference in the prediction accuracy before and after premutation is averaged over all the trees in the random forest, which defines the variable importance for X_j.

Random forests have two measurements of variable importance: Gini importance and permutation accuracy importance. Gini importance measures the sum of decreases in the Gini impurity when the node is split over all trees in the forest. Strobl observed that the permutation accuracy importance is more reliable than Gini importance in most situations (Strobl and Zeileis 2008). Another advantage of permutation importance is that it can be used for both continuous and categorical variables, whereas the Gini importance can produce biased result when the number of categories changes (Strobl et al. 2008).

The permutation variable importance for the t-th tree of variable X_j is calculated as

$$VI^{(t)}(X_j) = \frac{\sum_{i \in \bar{\mathfrak{B}}^{(t)}} I(\gamma_i = \hat{\gamma}_i^{(t)})}{|\bar{\mathfrak{B}}^{(t)}|} - \frac{\sum_{i \in \bar{\mathfrak{B}}^{(t)}} I(\gamma_i = \hat{\gamma}_{i,\pi_j}^{(t)})}{|\bar{\mathfrak{B}}^{(t)}|} \tag{2.1}$$

where $\bar{\mathfrak{B}}^{(t)}$ is the out of bag sample in tree t; t ranges from 1 to $ntree$ and $ntree$ is the number of trees in the forest; $\gamma_i = \hat{\gamma}_i^{(t)}$ and $\gamma_i = \hat{\gamma}_{i,\pi_j}^{(t)}$ represent the i-th observation's predicted class before and after permutation. The VIMP for X_j is the average of the score over all trees, which is the formula below

$$VI(X_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(X_j)}{ntree} \tag{2.2}$$

Since the $VI^{(t)}(X_j)$ are independent of one another in terms of different trees, the following test statistic has been proposed for testing the null hypothesis that the X_j's variable importance score is equal to 0. z_j is the z score for variable X_j; $\hat{\sigma}$ is the observed standard deviation of variable importance scores, and $VI(X_j)$ is the variable importance score for X_j under Breiman's definition.

$$z_j = \frac{VI(X_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}} \tag{2.3}$$

However, Strobl and Zeileis (2008) observed that the test statistics and p-values produced by this approach tend to be strongly anticonservative. It is inclined to increase type II error as the number of trees increases even if the null hypothesis is true.

2.2.3 Strobl's Conditional VIMP

Breiman's VIMP for variable X_j is calculated by permuting X_j independently of the other predictors, which is analogous to sampling from the marginal distribution of X_j. As a result, variables that are not associated with the outcome but are correlated with other important variables may still have high VIMP scores (Strobl et al. 2008). To overcome this shortcoming, one may sample from the conditional distribution of X_j conditioned on all of the other predictors denoted as X_{-j} rather than merely permuting X_j. Strobl proposed a method to sample from this conditional distribution that we call "Strobl's Conditional VIMP" (Strobl et al. 2008).

A description of the procedure is given below:

1. Before permuting X_j, calculate the out of bag prediction accuracy.

2. Conditioning on X_{-j}, identify the cutpoints that split the variable in a given tree and create a grid by bissecting the variable in each cutpoint.

3. X_j is permuted within the grid and OOB prediction accuracy is calculated after the

permutation.

4. VIMP of X_j within one tree is calculated by taking the difference of the OOB prediction accuracy before permutation and after permutation. The VIMP of X_j is calculated by averaging these importance score over all trees.

Compared with Breiman's VIMP, Strobl's conditional VIMP is less likely to assign high importance scores to spurious predictors that are correlated with other predictors. However it is not able to solve this issure entirely. It still tends to assign nonzero importance to such variables that are correlated with other strong predictors but not the outcome.

2.3 OPPERA study

2.3.1 Temporomandibular Disorder (TMD)

Temporomandibular disorder (TMD) is a painful disorder characterized by pain in the mastication muscles and temporomandibular joints. Although it is not life-threatening, patients with TMD may have constant pain in those regions as well as the head and neck muscles (Maixner et al. 2011a). Quality of life is negatively affected by this disease. TMD appears more often in females than males, with a prevalence of 6% in women versus 3% in men. A national survey suggested that 5% of adults in the U.S. have TMD (Isong et al. 2008).

TMD is a complex disease with multiple risk factors, and many of them are unknown (Cairns 2010). Possible risk factors include trauma and genetic variants, anatomical and psychosocial factors (Maixner et al. 2011a). Sociodemographic factors such as age, gender, and race, as well as clinical factors such as pain in other sites of the body are also the risk factors for TMD. Plesh et al. (2002) reported that TMD is more prevalent among African American women than Caucasian women. TMD is also related to pain amplification, which occurs when the nervous system enhances the response from the periphery

(e.g., muscle to brain) (Maixner et al. 2011a). Additionally, some psychological factors are associated with TMD. People with TMD are more likely to have depression and anxiety (Maixner et al. 2011a). Furthermore, genetic factors interact with environmental factors to affect the risk of developing TMD. Since TMD is a multifactorial and complex disease, the number of possible genes that correlated with TMD is large, such as monoamine oxidase A (MAOA) (Karayiorgou et al. 1999), glucocorticoid receptor (GR) (WuĬĹst et al. 2004), and D2 dopamine receptor (DRD2) (Lawford et al. 2003).

2.3.2 Overview of OPPERA

The "Orofacial Pain: Prospective Evaluation and Risk Assessment" (OPPERA) Study is a large-scale prospective cohort study to identify the genetic, psychosocial, autonomic, pain sensitivity and clinical factors for the development of TMD. Specifically, OPPERA aimed to determine if sociodemographic factors (such as age, gender, race etc.), elevated response to the noxious stimuli, and psychological factors are related to the first-onset of TMD. OPPERA also considered genetic factors for chronic TMD and first-onset TMD. OPPERA study has several cohorts and study designs, including a prospective cohort study, a baseline case-control study, and a matched case-control study. OPPERA recruited participants from four sites in the U.S., including The University of Maryland at Baltimore, MD, The University of Buffalo, NY, The University of North Carolina at Chapel Hill, NC and The University of Florida at Gainesville, FL. After enrollment, all the participants' information was collected via questionnaires, clinical examinations, and blood samples (Slade et al. 2011).

The OPPERA prospective cohort study recruited participants without TMD between 2006 to 2008. They were followed for a median of 2.8 years to see which participants developed first-onset TMD. Participants were recruited via advertisements, emails, flyers and word of mouth. 3,263 TMD-free participants were recruited between 18 and 44 years old .

Exclusion criteria included orthodontic treatment, pregnancy or nursing, history of injury or surgery on the face, and history of several serious medical conditions. At the baseline visit, potential enrollees signed the consent form and completed a series of questionnaires. Then they underwent a physical examination to confirm that they did not have TMD. Quantitative sensory testing to evaluate sensitivity to noxious stimuli was also performed. Furthermore, blood samples were collected and autonomic function was measured. The enrollees completed a quarterly health update (QHU) to evaluate the presence of pain in the face and jaw. Participants who reported orofacial pain on the QHU were instructed to return to the clinic for a follow-up clinical exam to evaluate the presence or absence of first-onset TMD (Slade et al. 2011). The baseline case-control study enrolled patients between 2006 and 2013 with around 1000 people with TMD as cases and 3200 TMD-free controls. TMD cases were determined using Research Diagnostic Criteria for Temporomandibular Disorder (RDC/TMD), including more than 4 days of facial pain during the previous 30 days and pain motivated by jaw movement and temporomandibular joints during the examination.

2.3.3 Studies of putative risk factors and TMD

The association with sociodemographic and socioeconomic risk factors and TMD were studied in both baseline case-control study and the community-based prospective cohort study. In the study of 1633 controls versus 185 cases among participants aged from 18 to 44, positive association was identified with patients from greater age; the odds ratio of female and male with TMD was four; White non-hispanic participants was less likely to have TMD than the other races. While whether participants were born in the USA was not important to chronic TMD (Slade et al. 2011). Consistent with case-control study, the prospective cohort study showed that patients with greater age group were more likely to have TMD than younger people. The association of gender with first-onset TMD was

marginally significant. Black American had large odds than white people for TMD. U.S residency for lifetime are strong predictor and people spend their life in the USA had higher odds than the others (Slade et al. 2013). The association of TMD and socioecomonic characteristics illustrated that English as the first spoken language had higher odds of chronic TMD than those with other first spoken languages. Never married people were more likely to have chronic TMD than married ones (Slade et al. 2011). In prospective cohort study, higher satisfaction with material standards in life contributes less to first-onset TMD (Slade et al. 2013).

Many case-control studies illustrated that people with chronic pain have higher level of psychological maladjustment compared with pain-free controls (Dworkin et al. 1990), and TMD is one of them. Chronic TMD cases showed greater level of global measures of psychological functions, affective distress and stress, somatic awareness and coping/catastrophizing (Fillingim et al. 2011). Similarly researchers revealed the association with increased risk of first-onset TMD and psychological factors, such as somatic symptoms, psychosockal distress and affected distress. But coping is not the predictors for first-onset TMD (Slade et al. 2007)

In the previous first-onset studies, it is identified that TMD were associated with some baseline self-reported measurements, including oral parafunctions, number of nonspecific orofacial symptoms and other clinical risk factors (Ohrbach et al. 2013), pain sensitivity and heart rate at rest in pain sensitivity risk factors (Greenspan et al. 2013), global psychological and somatic symptoms as well as passive pain as the most important rpsychological factors (Fillingim et al. 2013).

2.3.4 Genetic Associations with TMD in OPPERA

TMD is a complex disease with multiple risk factors. In particular, genetic factors are believed to contribute to the risk of TMD. There are several markers that were associated with TMD in previous studies, such as the T102C SNP of serotonin receptor HTR2A (Mutlu et al. 2004), A218C SNP in TPH1 gene (Etoz et al. 2008), and haplotypes of beta-2-adrenergic receptor (ADRB2) (Diatchenko et al. 2006a). One goal of OPPERA is to identify genes associated with chronic TMD. A total of 358 candidate genes involved with pain processing were selected, and each OPPERA participant was genotyped at a series of SNP's in each gene. (Smith et al. 2011). Blood samples were collected from study participants and genotyped. The analytic data sets have 1,961 observations (including 348 chronic TMD cases and 1,612 TMD-free controls) who were genotyped for 2,657 SNPs. Logistic regression was performing using the PLINK software to see if each SNP was associated with TMD case status after adjusting for other covariates such as age, gender, and race.

The SNP's were classified into two tiers. The first tier included 23 candidate genes that were considered to be high priority, whereas the second tier included all 358 genes. Bonferroni correction was used to adjust for multiple comparisons. In the analysis of the Tier 2 SNP set, 9 SNPs in 6 genes were correlated with TMD case status. Specifically, 3 SNPs were in the long intron of NR3C1 gene, which is a gluccorticoid receptor gene; 2 SNPs were in located in the CAMK4 gene, the calcium dependent protein kinase 4 gene; and the remaining 4 SNPs were located in CHRM2 (muscarinic cholinergic receptor 2), HTR2A (serotonin 2A receptor), IFRD1 (the first intron of interferon-related developmental regulator 1) and GRK5 (the second intron of G protein-coupled receptor kinase 5) genes, respectively. Among the Tier 1 SNPs, 2 SNPs in interleukin 10 (IL10) gene, 3 SNPs in an adrenergic receptor genes (ADRA2C), and one SNP in the delta opioid receptor (OPRD1) gene, showed significant associations with TMD (Smith et al. 2011).

CHAPTER 3: A PERMUTATION-BASED GENETIC ASSOCIATION TEST FOR SECONDARY PHENOTYPES

3.1 Introduction

In a cohort study, participants are followed over a period of time to evaluate the association between an outcome of interest (such as a disease) and exposures (risk factors). This study design allows one to measure exposures before the disease occurs and estimate the incidence rate, which is an attractive design for studies in common diseases. However, cohort studies take a long time to complete and are relative expensive, which is problematic in genetic studies. Moreover, the sample size would be prohibitively large for rare diseases in order to have sufficient power. Compared with cohort studies, case-control studies are much more cost-effective. They are attractive alternatives where researchers select participants with disease (cases) and without the disease (controls). Although case-control studies have a greater potential for confounding and do not allow one to estimate the disease's incidence rate, they are generally much cheaper and easier to perform than cohort studies. In particular, genome-wide association studies (GWAS) typically use case-control design.

GWA studies are expensive in both time and cost, and researchers want to collect as much information as possible to maxmize the return. In addition to identifying single-nucleotide polymorphisms (SNP) genotype's associated with the disease of interest, secondary phenotypes, which are associated with primary outcome are also collected and studied. For example, in a study of TMD, one may be interested in SNP's association with pain sensitivity in addition to SNP's association with TMD case-control status. TMD patients tend to be more sensitive to pain than TMD-free controls, so the secondary outcome

pain density might be correlated with TMD case status. Similary, the association between body mass index (BMI) and SNPs were studied to understand the etiology of type II diabetes (Frayling et al. 2007). In the GWA study of lung cancer, the correlation of secondary outcome smoking and genetic variants were investigated to reveal useful information in lung cancer (Villeneuve and Mao 1993) (Hung et al. 2008). Moreover, secondary phenotype provide more information about disease and are useful in gene mapping as well (He et al. 2011).

The secondary phenotype studies should be analyzed carefully to avoid problematic or even misleading results. The common approaches to analyze the association of secondary outcome and SNPs in case-control studies are standard regression methods by using cases only, controls only, samples with both cases and controls, or joint analysis of cases and controls adjusting for disease status. However, the above methods led to spurious association of secondary outcome and genetic variants because the samples from both cases and controls are not representative of random samples from population. Hence none of them are reliable (Monsees et al. 2009, Lin and Zeng 2009).

Data from case-control studies are not a representative sample of the population since cases are overrepresented. it has been shown that logistic regression produce unbiased coefficient estimates for estimating the risk of case status in a case-control study. (Prentice and Pyke 1979) However, there is no guarantee that coefficient estimates are unbiased if the outcome is an secondary phenotype rather than case status. Indeed, logistic regression without adjustment for the study design will produce biased estimates (Monsees et al. 2009).

A number of statistical methods have been developed for the remedies. Richardson et al. (2007) recommended inverse probability weighting (IPW), whereby cases and controls are weighted such that the total weights of cases are comparable to the proportion

of cases in the general population. However, standard logistic regression models will produce incorrect estimates of the standard error after applying such weights. Monsees et al. (2009) demonstrated that one may calculate the standard error correctly using a sandwich estimator of the variance. The estimates by IPW method can be unstable when the secondary phenotype is strongly associated with case status. For example, in genetic studies of TMD, one may wish to find genetic factors associated with the severity of orofacial pain. Essentially all TMD cases will report some level of orofacial pain, but the majority of controls won't. When standard IPW regression is applied in this situation, tests of no association between a given SNP and pain severity produce widely inflated type I errors. In this situation, a more robust method is needed , which will help to identify genetic risk factors for TMD, or be used as a tool for future similar situations in other GWA studies. Wang and Shete (2011) proposed a bootstrap-based method for the estimation of the odds ratio for a binary secondary outcome. However, it cannot be applied to a continuous outcome. Lin and Zeng (2009) proposed a method based on maximum likelihood that can be applied to both continuous and categorical intermediate phenotypes. This method is only unbiased when the genotypes are not related to the primary outcome (Li et al. 2010).

In this chapter, we propose a permutation-based IPW method to analyze the association between a genetic marker and secondary phenotypes in a case-control study where the secondary outcome is correlated with case status. The proposed method has the advantage of being completely non-parametric, so it will produce valid p-values even when the assumptions of parametric models are violated (which is a common occurence when examining secondary phenotypes in the OPPERA study). We showed that the proposed method has comparable power to existing parametric methods and it avoided the risk of elevated type I error when the model assumptions are violated. We further applied the method to the OPPERA data, and found two SNPs highly associated with clinical pain density.

3.2 Method

3.2.1 Permutation-based IPW

The assumption of conventional IPW method is violated when secondary phenotype is strongly correlated with cases, which will lead to anti-conservative p-values, and hence result in inflated type I error. In order to have valid p-values, we proposed a novel permutation-based IPW method to evaluate the association between genetic markers and secondary phenotypes in a case-control study when both are correlated with disease status. The proposed method was motivated by the idea that when a given SNP (X_j) is associated with secondary phenotype (Y), it is expected that the counts of minor alleles for the SNP genotype is highly correlated with secondary outcome. When the Y is permuted as Y^*, the inherent association between SNP and secondary outcome is destroyed. We expected to see that the association between the permuted secondary outcome and SNP genotype would be smaller than that before permution. So, we may test the null hypothesis of no association between secondary outcome and a given SNP genotype (denoted as SNP_j) by comparing the association of SNP and secondary outcome before and after permutation.

Suppose the number of participants with and without primary disease are n_1 and n_0 respectively in a case-control study. Let X_{ij} be the number of copies of the minor alleles for SNP j subject i; let Y_i be the secondary phenotype for subject i. Assume there are N SNPs. We assumed that the disease prevalence is s and hence the probability of controls is $1 - s$ in the population. Then the weight of cases in the sample would be comparable to that in general population if we assigned a weight of 1 to controls. Thus, the weight of cases (defined as Wt_1) was calculated by the following formula:

$$Wt_1 = \frac{s \times n_0}{(1-s) \times n_1} \tag{3.4}$$

Nonparametric method permutation was utilized to estimate the p-value of testing the

null hypothesis of no association between X_j and Y. Y was permutated B times as Y_1^*, Y_2^*, Y_3^*, ..., Y_B^*. The weighted correlations between X_j and each of Y, Y_1^*, Y_2^*, Y_3^*, ..., Y_B^* were calculated for every SNP_j with weights derived from IPW weights given above. The p-values (denoted as p_j) for testing the null hypothesis of no association between SNP_j and secondary outcome was estimated by dividing the number of times that more extreme correlations are observed across all the permuted correlations (Bair 2013). The function of estimating p-value is shown below.

$$p_j = \frac{1}{NB} \sum_{i=1}^{N} \sum_{k=1}^{B} I(\left|\mathrm{R}_{l,k}^*\right| \geq |\mathrm{R}_j|), \tag{3.5}$$

where R_j is the weighted correlation between X_j and Y, and $\mathrm{R}_{l,k}^*$ is the weighted correlation between X_l and permuted outcome Y_k^*. $I(x)$ is the indicator function. It is equal to 1 when the condition is true, and 0 when the condition is false. The p-value was calculated by pooling the weighted correlation across all the candidate genes and permuted secondary outcome which is counting the number of absolute value of $R_{l,k}^*$ greater than absolute value of R_j divided by NB. Note that this procedure may be used even if Y_i is binary, since the correlation between X_i and Y_i is proportional to the Armitrage trend test statistic. This method assumes that the distribution of $R_{l,k}^*$ does not depend on j.

Now suppose one wishes to evaluate the association between an allele count X_i and a secondary phenotype Y after controlling for covariates $Z_1, Z_2, \ldots, Z_k$. The above procedure can be modified as follows:

1. Regress X_i on the Z_is by weighted least square regression where the weights were calculated in the above method. Find out the residuals vector $X_i^{'}$ from the resulting model.

2. Similarly, Regress Y on the Z_is by weighted least square regression and get the residuals vector $Y^{'}$ from the resulting model.

3. Apply the permutation test procedure above using $X_i^{'}$ and $Y^{'}$ in place of X_i and Y.

This procedure is useful when one wishes to adjust for demographic covariates or eigenvectors corresponding to population stratification such as race or ancestry (Price et al. 2006).

The above procedure requires one to regress X_j on the covariates for each SNP. Thus, a naive application of this precedure would require the computation of millions of regression models, which would be computationally expensive. By noting that each regression model has exactly the same covariates and X_j is the only difference, we proposed a method that can significantly reduce the computing time. Let Z be the model matrix for a given regression model, where each column of Z is an individual covariate. Let W be the $N \times N$ diagonal matrix of weights, and the weights on the diagonal were 1 for controls and Wt_1 for cases. The weighted least squares regression coefficients were given by

$$\hat{\beta}_w = (Z^T W Z)^{-1} Z^T W X_j \tag{3.6}$$

The estimated values of X_j were calculated as

$$\hat{X_j} = Z(Z^T W Z)^{-1} Z^T W X_j \tag{3.7}$$

Let $P = Z(Z^T W Z)^{-1} Z^T W$, and the above equation becomes

$$\hat{X_j} = P X_j \tag{3.8}$$

Noting that P depends on Z not X_j, one may calculate and store matrix P. Then the residuals of the weighted regression model to predict X_j on Z (denoted as $\hat{e}_{Xj}$) was calculated by

$$\hat{e}_{Xj} = X_j - P X_j \tag{3.9}$$

Hence, the residual matrix for X was

$$\hat{e}_X = X - PX \tag{3.10}$$

Similarly, the residuals of Y on Z (denoted as $\hat{e}_Y$) was using the following formula,

$$\hat{e}_Y = Y - PY \tag{3.11}$$

This requires only a single matrix multiplication rather than recomputing the entire regression model for each SNP, and it is likely to significantly reduce the computation needed for this procedure.

3.2.2 Type I error simulation

The above procedures were conducted to produce valid p-values whereas the conventional IPW method produced inflated p-values. Simulations were performed to evaluate the type I errors of the proposed method as well as a conventional IPW method. In the first simulation, a data set with 100 "subjects" and 10000 "SNPs" were simulated. The first 50 "subjects" were designated as cases, and the remaining "subjects" were designated as controls. The secondary outcome variable was simulated as integers between 0 and 10. For cases, the secondary outcome variable was generated under a uniform distribution with the integers between 1 and 10. To generate the secondary outcome for controls, first of all a proportion p was generated under a (continuous) uniform distribution on $(0.01, 0.1)$. Then for each control in the sample, the secondary outcome variable was defined to be 0 with probability $1 - p$ and an integer generated uniformly between 1 and 10 with probability p. This simulation scenario is motivated by secondary pain phenotypes in the OPPERA

study, where all cases report some level of pain and most controls report no pain. The minor allele frequency (MAF) of each SNP was randomly generated from a uniform distribution ranging on $(0.05, 0.1)$. After generating the MAF, the number of minor alleles for each subject at each SNP was generated under a binomial distribution with two trials with the probability of success equal to MAF.

The second simulation was performed in the scenario where population stratification was present in the simulated SNPs. The data was generated by a Balding-Nichols model similar to the simulations in Price et al. (2006). This simulated data set had 1,000 "subjects" (with 500 controls and 500 cases) and 10,000 SNPs. The secondary outcome variable was generated by the same procedure as those in the first simulation. To generate SNPs, an ancestral allele frequency p was generated for each SNP under a uniform distribution on $(0.1, 0.9)$. We assumed that two subpopulations exist in the sample. The allele frequency for each subpopulation (denoted as p_i) was generated using Balding-Nichols' model, as the formula below:

$$p_i \sim \text{Beta}\left(\frac{p(1-F_{ST})}{F_{ST}}, \frac{(1-p)(1-F_{ST})}{F_{ST}}\right) \tag{3.12}$$

where p_i is the allele frequency for population i, where $i = 1, 2$. The coefficient $F_{ST} = 0.01$ represents Wright's coefficient of inbreeding, which was chosen based on the inbreeding coefficient observed in different populations of Europen ancestry (Pritchard and Donnelly 2001). In the first 100 SNPs, the ancenstral allele frequencies were chosen to be 0.2 and 0.8 for population 1 and 2 respectively, to model the fact that a small number of SNPs are expected to have larger differences between the two populations. This simulation scenario is analogous to the model used for the simulations in (Price et al. 2006). It was assumed that there were 300 cases and 200 controls in population 1 and the remaining simulated subjects were in population 2. For each SNP, the count of minor alleles for each subject was generated under a binomial distribution with two trials and probability of success p_i,

as described previously.

In both simulation scenarios, the secondary phenotype was permuted 10 times. The p-value for testing the null hypothesis of no association between a SNP and the secondary phenotype was calculated by proposed permutation-based IPW method as described in Section 3.2.1. The p-value for conventional IPW regression was obtained using the "geepack" R package as described in Monsees et al. (2009). The prevalance of the disease was assumed to be 5% in the general population and was utilized in calculating the weights for the cases. In the second type I error simulation, the first 10 eigenvectors were taken into account as covariates adjusting for population stratification. No covariates were included for the first simulation since only one "population" were simulated in the first data set. The p-values were generated by proposed methods and conventional IPW method to test the null hypothesis. Note that there's no association between the SNPs and the outcome variable in both simulated scenarios, all "significant" associations are necessarily false positives. Under null hypothesis, all the p-values should be therefore uniformly distributed between 0 and 1. When a method produces excessive number of small p-values, that indicates that the method produces inflated type I errors and would have spurious association with secondary outcome. Q-Q plots were utilized to present the observed p-values versus the expected (uniform) distribution of the p-values for conventional IPW method and the permutation-based IPW method. Lines were added to the plots showing the expected distribution of the p-values, as well as the significance thresholds corresponding to false discovery rates (FDR's) (Benjamini and Hochberg 1995) of 0.2, 0.1, 0.05 respectively. When a SNP point on the plot is above the line corresponding to an FDR of 0.05, that implies that the estimated false discovery rate would be less than 0.05 if that SNP (and all other SNPs above the line) were called "'significant'.

3.2.3 Power Simulation

We hypothesized that this method will have comparable power to conventional IPW regression. In order to evaluate the power of the proposed method, we generated 10,000 SNPs for 2,000 subjects using a simulation similar to the simulation proposed by Lin and Zeng (2009). Let Z be an indicator for case status; let Y_i denote a quantitative secondary outcome, and let X_i denote the number of minor alleles of SNP i. Then each X_i was generated as a binomial random variable with two trials and probability of success p, where p is a tunable parameter. Then Y_i was generated as follows:

$$Y_i = \beta_0 + \beta_1 X + \epsilon_i \tag{3.13}$$

where ϵ_i has a normal distribution with mean 0 and variance 1. In this simulation, we let $\beta_0 = 0$, and β_1 was a tunable parameter.

To model the dependence of case status on both X_i and Y_i, the probability of case status was defined to be

$$P(Z = 1|X_i, Y_i) = \frac{e^{g_0+g_1X_i+g_2Y_i}}{1 + e^{g_0+g_1X_i+g_2Y_i}} \tag{3.14}$$

where g_1 and g_2 were equal to $\log 2$; and g_0 was defined to be

$$g_0 = \log \frac{\phi}{1-\phi} - g_1\bar{X} - g_2\bar{Y} \tag{3.15}$$

In the formula above, $\bar{X}$ and $\bar{Y}$ are the mean value of X_i and Y_i, and we let $\phi = 0.1$. (Under this model, the prevalence of the disease will be approximately ϕ.)

For each simulated observation, the probability of case status was calculated based on (3.14) and the observation was randomly assigned to be a case or a control based on this probability. The process was continued until 1000 controls and 1000 cases were generated. (Extra controls were discarded once 1000 controls had been generated.) After a

complete data set was generated, both conventional IPW regression and our permutation-based IPW method were applied to the simulated data set. The null hypothesis of no association between the SNP and the outcome was rejected if the p-value was less than 0.05. These calculations were repeated 10000 times for each choice of p and β_1. The power of each method was estimated to be the number of times the method produced a p-value of less than 0.05 divided by the number of repeated times (which is 10000).

3.2.4 Data Application

We applied our method to data from the "Orofacial Pain: Prospective Evaluation and Risk Assessment" (OPPERA) baseline case-control study. OPPERA study is a large scale prospective cohort study in finding out the etiology and putative risk factors of chronic and onset of temporomandibular disorder (TMD). In this study, people with TMD were consideredd as cases, while people without TMD were denoted as controls. There are compelling evidence that the genetic factors play an important role in the etiology of TMD. Genetic inheritance contributes to about 27% to the variation of TMD pain in a recent twin study (Plesh et al. 2012). Numerous studies found out associations between TMD and genetic factors. While these studies are informative, one may wish to identify the genetic variants that associated with secondary phenotype, such as widespread body pain and clinical pain. TMD patients with greater widespread body pain or clinical pain might represent a homogenous cluster of TMD patients as treating all the TMD patients homogenously would fail to detect genetic markers associated with the most severe forms of TMD. Each patients in this study filled out a set of questionnaires at screening, and underwent a series of clinical examinations to identify risk factors of TMD. A blood sample was obtained from each study participants and and stored in 5 ml EDTA-containing polyethylene vacutainers at $-80\,^{\circ}\mathrm{C}$. Then the samples were genotyped using Omni2.5 Bead Chip Illumina Platform as part of a genome-wide association study. Genetic data

was available for both cases and controls.

The study enrolled a total of 3263 TMD-free controls and 186 chronic TMD cases. Genetic data on 2924 SNPs was available for 3050 of these subjects. The secondary phenotype for this analysis was the pain density, a quantitative trait calculated as the average self-reported orofacial pain rating (on a 0-100 scale) multiplied by the self-reported percentage of the day with pain. We included participants with non-missing pain density and all SNPs with a MAF greater than 0.02. The final dataset consists of 3001 patients (with 166 TMD cases and 2835 TMD-free controls) and 2864 SNPs. Using the method described above, principal component analysis was performed on the SNP matrix, and the first 6 components were included in the model as covariates controlling for population stratification. Gender and dummy variables for OPPERA study sites were also included as covariates. The permutation-based IPW procedure was applied to calculate p-values for each SNP as described in Section 3.2.1. The p-values of testing null hypothesis of no association between pain intensity and a SNP were visualized in Q-Q plot, where under the null hypothesis the p-values were expected to follow uniform distribution.

3.3 Results

3.3.1 Type I Error

The Q-Q plots of the observed p-values versus the expected p-values under null hypothesis of no association between secondary phenotype and a given SNP by permutation-based method and conventional IPW method were shown in Figures 3.1 and 3.2 respectively. They represent the scenario in simulation 1 where only one population were taken into consideration. Each point on the plot represents a simulated "SNP". Under the null hypothesis, we expected to see that the expected p-values were nearly identical to the observed p-values since none of the simulated "SNP" were associated with simulated "secondary phenotype". In Figure 3.1, almost all the SNPs are on the line where the expected

p-values are equal to the observed p-values, and none of the SNP point is above any of the lines corresponding to FDR 0.2, 0.1 nor 0.05. This implies that we do not reject the null hypothesis of no correlation between a "SNP" and secondary phenotype, which is what we expected to see in the simulation. Type I error is well preserved by our nonparametric IPW method in this scenario. While Figure 3.2 is the Q-Q plot of p-values from the same simulated data as those from Figure 3.1, but using conventional IPW method under null hypothesis. We can see that a lot of SNP points' with obseved p-values largely different from the expected p-values, even above the lines with FDR 0.05. We rejected the null hypothesis and those SNP points above the line where $FDR = 0.05$ are significantly associated with secondary outcome by looking at the plot. However, this is contradictory to our simulated data where none of the SNP is correlated with secondary outcome. This implies that conventional IPW method producedd unstable p-values and would inflate type I errors. We cannot use this method in the analysis of real data which are similar to the simulated data, since it would have false positive results and lead to spurious association between them.

Figure 3.3 and 3.4 showed the p-values for both permutation-based IPW and conventional IPW method under null hypothesis when population stratification were considered as illustrated in simulation 2. Similar as the result in simulation 1, almost all the "SNP" points were on the anti-diagonal line where expected p-values were equal to the observed p-values. Only two points were on or above the lines with FDR 0.2 and 0.1. None of the points were above the line with FDR = 0.05. Our method do not have inflate p-values in scenario 2 and type I errors were well conservative. Compared with nonparametric IPW method, conventional IPW method shown in Figure 3.4 had a trend of inflated type I errors, meaning that the observed p-values were greater than expected p-values for a number of simulated "SNPs", which in fact did not associated with "secondary phenotype" in the second simulation.

In both simulation scenarios, the p-values produced by conventional IPW regression deviated significantly from the expected p-values, indicating that conventional IPW regression is producing anticonservative p-values in these simulation scenarios. In contrast, the distribution of the p-values produced by permutation-based IPW procedure are nearly identical to the expected null distribution. Our method is likely to find out true association of a SNP and secondary outcome rather than spurious associations which are likely to be produced by conventional IPW method.

3.3.2 Statistical Power

Table 3.1 shows the estimated power of conventional IPW regression and permutation-based IPW approach for various values of MAF and β_1. The powers by both methods were very similar, indicating that our method has comparable power to conventional IPW regression when true association exists.

3.3.3 Data application

The above results showed that our method has conservative type I errors and comparable power to conventional IPW method. The application of the method to OPPERA data would help us to identify genetic markers associated with secondary phenotype (which is the severity of TMD pain). Figure 3.5 is the Q-Q plot of the observed p-values versus the expected p-values under the null hypothesis of no association between a given SNP and pain density in the OPPERA study. Most of the data points were on the anti-diagonal line where expected p-values were identical to the observed p-values, implying that those SNPs were not associated with pain density. Two SNPs show evidence of an association with the pain density phenotype, including one SNP that shows a significant association after adjusting for multiple comparison.

3.4 Discussion

IPW regression is a useful tool for evaluating the association between genetic parkers and secondary phenotypes in GWAS studies. However, it can produce unreliable estimates when the assumptions of the model, where the secondary phenotype is strongly associated with disease case-control status are violated. In particular, conventional IPW regression performs poorly when evaluating the association between a genetic marker and a secondary phenotype that is equal to 0 for the majority of the controls. IPW regression gives lower weight to cases to account for the fact that they are overrepresented in a case-control study. Thus, if the secondary phenotype is equal to 0 for the majority of controls, nearly all of the variance in the secondary phenotype outcome variable occurs among the data with low weight, resulting in unstable regression estimates.

This situation is common in studies of chronic pain such as OPPERA. Suppose we wish to identify SNPs associated with the severity of orofacial pain. All TMD cases have some level of orofacial pain, but the majority of controls report no pain in the orofacial region. If conventional IPW regression is applied to OPPERA (or similar data sets), it produces a large number of anticonservative p-values.

In addition to conventional IPW method, other methods proposed for the analysis of secondary phenotype and a genetic variants generally use parametric regression models (Lin and Zeng 2009, He et al. 2011, Wang and Shete 2011). There's a need of using nonparametric methods rather than parametric ones where the assumptions are violated, to produce robust estimates. Our proposed permutation-based IPW procedure overcomes this shortcoming of conventional IPW regression. Our simulations and real data results indicated that this procedure avoid the problem of inflated type I error when the assumptions of the conventional IPW regression model are violated without sacrificing power. The procedure identified two novel SNPs associated with clinical pain in the OPPERA study. To our knowledge, our method is the first purely nonparametric hypothesis test to evaluate

the association between a risk factor and a secondary phenotype.

The proposed method is easy to implement using common statistical software and is relatively fast. Fewer permutations are required when the number of SNPs is large, so the method can be applied to GWAS data sets with millions of markers. Thus, the proposed method is an attractive alternative to conventional IPW regression, particularly for secondary phenotypes with few nonzero values among controls.

Our application to identify SNP's associated with pain density in TMD case-control studies is also novel. Most previous GWA studies in OPPERA focused on the TMD disease status, which was usually considered as the outcome of interest. Such studies do not consider the fact that the severity of the pain may vary greatly among TMD patients nor the fact that individuals who do not meet the diagnostic criteria for TMD may still experience some levels of pain. Although genetic association of TMD related secondary phenotype (such as pressure pain sensitivity and non-specific orofacial symptoms) were analyzed in previous studies, our analysis was focusing on pain density and based on candidate genes. This analysis is the first GWA study to identify the association between genetic markers and clinical pain severity, such as pain density. This analysis will lead to the new insights on the genetic risk factors for TMD and chronic pain more generally. In a recent study, Bair et al. (2016) found that TMD consists of at least three homogeneous subgroups or clusters. Some clusters are associated with more severe clinical pain measures. It is believed that patients in different clusters will respond differently to treatments. Identifying genetic markers associated with more severe forms of TMD will hep explaining why some patients have more severe TMD symptoms and may lead to novel treatment plans.

This methodology has some limitations. It only produces p-values. It cannot be used to calculate odds ratios (or any other measure of the strength of an association) or associated confidence intervals. While the variance of effet size measures produced by conventional

IPW regression are oftern anticonservative when a secondary phenotype is strongly associated with case status, resulting in confidence intervals that are too narrow. Future analysis with alternative approach is needed to address these questions.

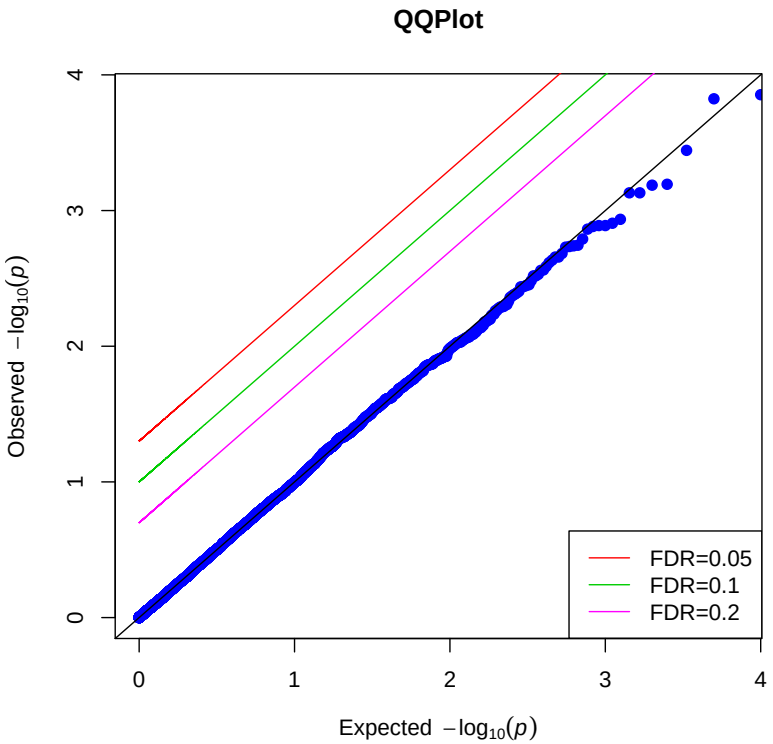

Figure 3.1: QQ plot of the p-values produced by the permutation-based IPW method for the first simulation scenario

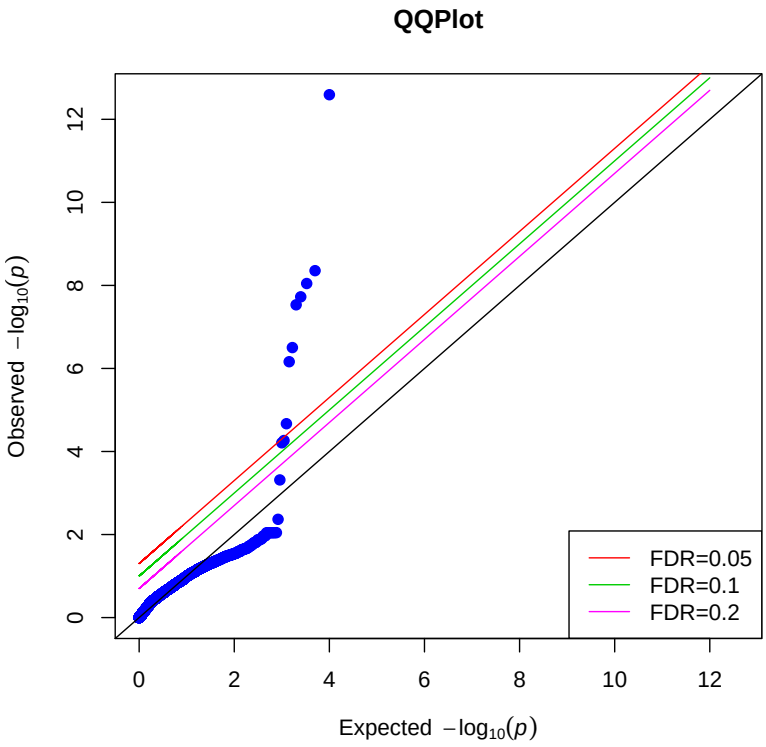

Figure 3.2: QQ plot of the p-values produced by conventional IPW regression for the first simulation scenario

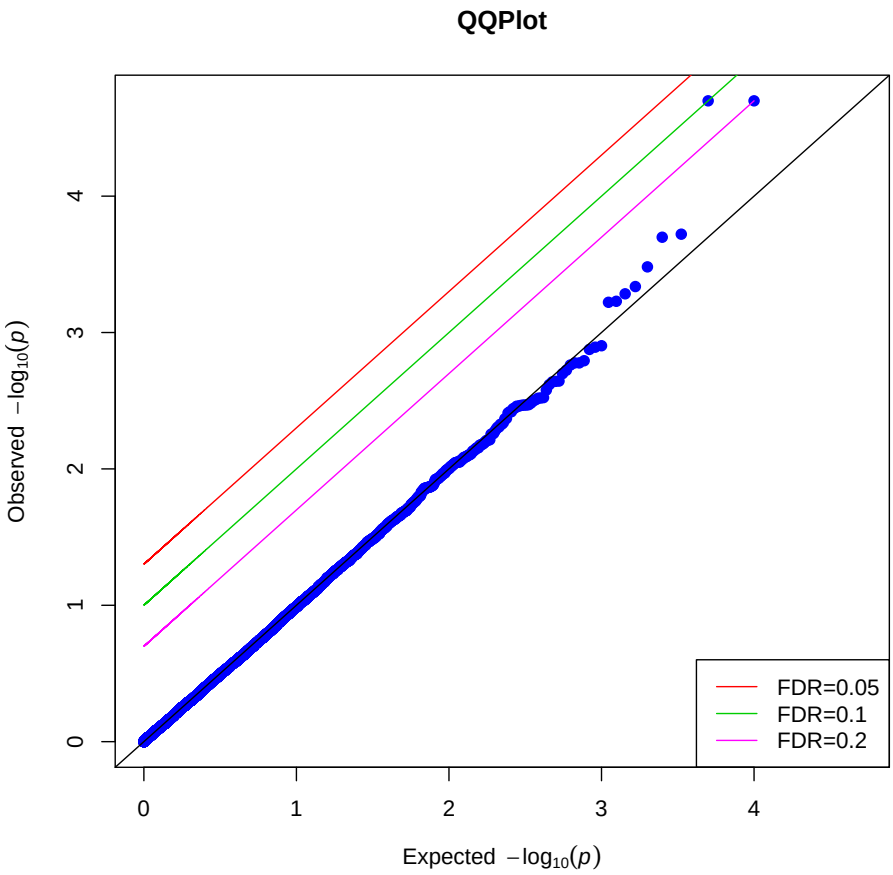

Figure 3.3: QQ plot of the p-values produced by the permutation-based IPW method for the second simulation scenario

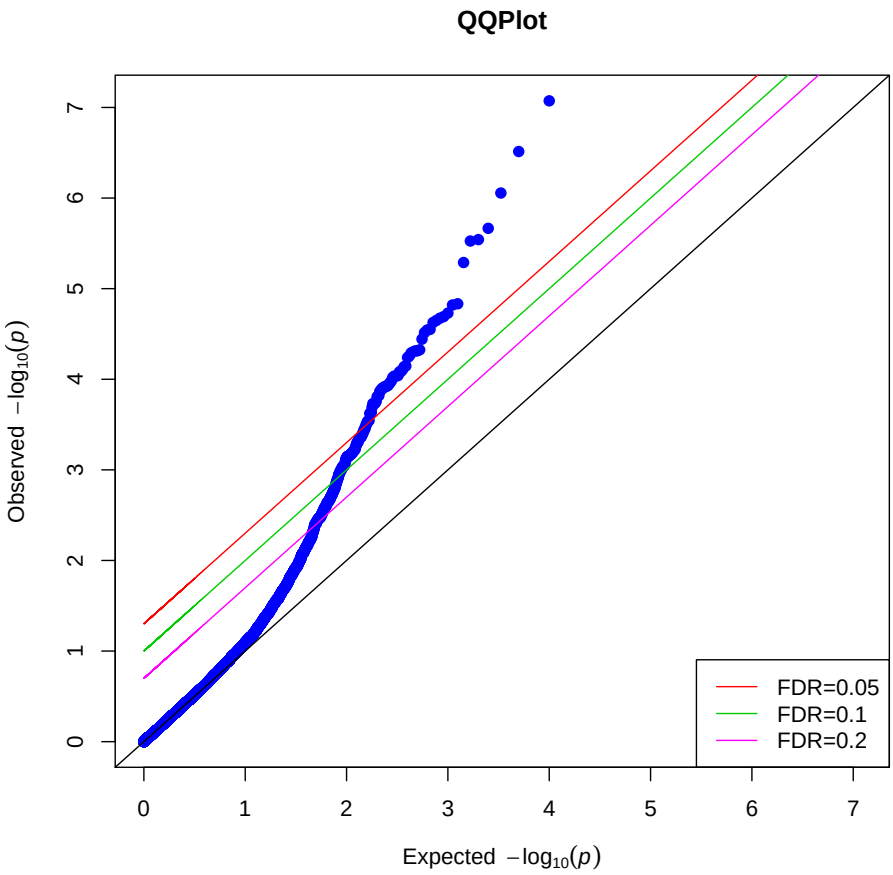

Figure 3.4: QQ plot of the p-values produced by conventional IPW regression for the second simulation scenario

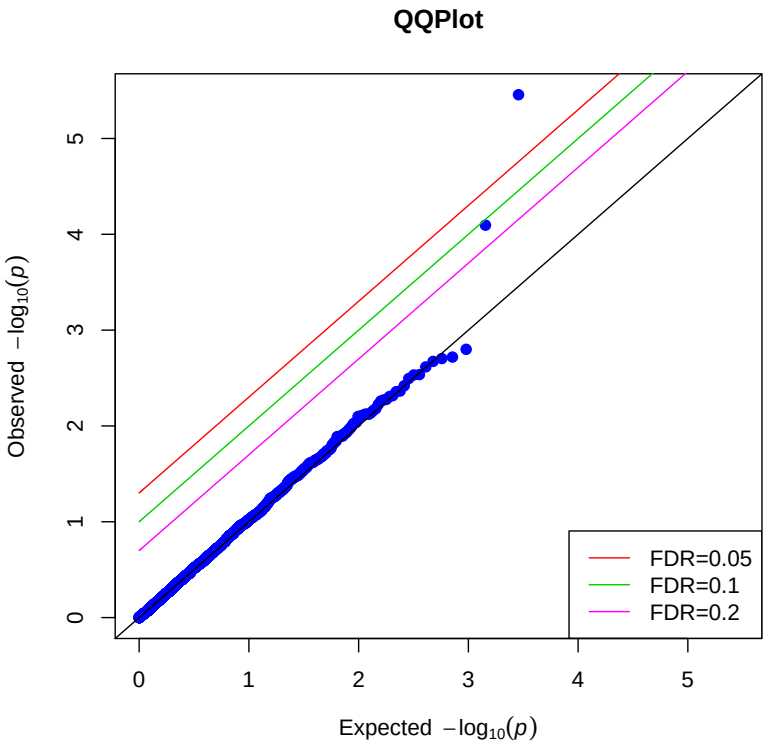

Figure 3.5: QQ plot of the p-values for testing the null hypothesis of no association between each SNP and pain density in the OPPERA study

Table 3.1: Power simulation

β_1	MAF	Power using permutation-based IPW	Power using conventional IPW regression
-0.12	0.06	0.2295	0.3302
-0.42	0.3	1	1
-0.17	0.3	0.9782	0.98
-0.16	0.3	0.9633	0.9633
-0.15	0.3	0.9407	0.9462
-0.13	0.3	0.8631	0.8723
-0.12	0.3	0.7956	0.8079
-0.12	0.4	0.8581	0.8620
-0.012	0.3	0.056	0.0599
0.12	0.3	0.8137	0.8256
0.14	0.3	0.9181	0.9223
0.18	0.3	0.9899	0.9908
0.2	0.3	0.9976	0.9977
0.2	0.4	0.9994	0.9994
0.4	0.4	1	1

CHAPTER 4: CONDITIONAL VARIABLE IMPORTANCE TEST IN RANDOM FORESTS

4.1 Introduction

Random forests are an ensemble learning method proposed by Breiman in 2001. The method combines bagging, classification and regression trees (CART). It performs very well compared with the other method, and it is widely used in genetics (Goldstein et al. 2010), ecology (Cutler et al. 2007), physics, bioinformatics and other fields. This tool is able to solve problems related to prediction even with nonlinearity and complex interactions.

Random forests combine conventional decision trees in the following ways. A set of bootstrap samples (which are sample with replacement) are selected as training sets to create trees in the forest. Those left out of the samples are called out-of-bag (OOB) data, which will further be utilized to get the prediction or classification error and calculate the variable importance. Within each resulting tree, a subset of variables are selected at each node, and they are splitted within the chosen variables (rather than splitting among all variables in conventional decision trees). This process is repeated until the tree is fully grown without pruning. Random forest predictors are obtained by averaging over all trees. This strategy decreases the correlation among trees and thus allows one to reduce the variance in a single decision tree by averaging (Archer and Kimes 2008). It also increases the prediction accuracy based on the ensemble method compared with single classification tree.

Random forests have many excellent features. It can be utilized when the number of

variables are large while the number of observations are small, known as a "small n large p" problem because each split on a node is based on a subset of variables rather than all the variables. Thousands of variables can be taken care of in the model without deleting any of them. It can be applied to both categorical, continuous outcomes and survival data. Random forests are able to handle highly correlated predictors and account for arbitrary interactions. They are also able to model nonlinear associations between predictors and the outcome. In general they will not overfit the data. Even though the results are not easy to interprate as those of classification tree does, varaible importance in random forest helps to solve the issue, as discussed below (Díaz-Uriarte and De Andres 2006). Missing data can be imputed through random forest (`https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`).

Variable importance can be obtained by random forests. The advantage is that it contains both the predictor's impact and the interactions with the others. Intuitively, a variable is "important" if the model's predictive accuracy decreases when the variable is removed from the model or measured with error. The naive way of vcalculating ariable importance is to count the number of times the variable is selected in the trees in tree-based methods. Gini importance and permutation accuracy importance are the two most common and more advanced measures of variables importance in random forests and usually they are very consistent. Gini importance measures the sum of decreases in the Gini impurity when the node is split over all trees in the forest. The rationale for premutation accuracy importance by Breiman is that the original correlation of a predictor variable X_j and outcome Y is broken by randomly permuting X_j. The prediction accuracy by permuted X_j and the other predictors decreases dramatically if X_j is highly associated with outcome. Hence the difference of prediction accuracy is the variable importance (Strobl et al. 2007). Strobl observed that the permutation accuracy importance is more reliable than Gini importance in most situations (Strobl and Zeileis 2008). Another advantage of

permutation importance is that it can be used for both continuous and categorical variables, whereas the Gini importance produces biased result when the number of categories changes (Strobl et al. 2008).

Breiman's VIMP can be used to identify important variables in the model, but it has certain shortcomings. It tends to assign high importance to variables that are correlated with those which are highly associated with the outcome even if these variables are not associated with the outcome (Strobl et al. 2008). Breiman also proposed a statistical test for variable importance with the null hypothesis that the variable importance score for a given variable is equal to 0 based on z-scorse which is the scale of variable importance (Breiman and Cutler 2008). But The z-score tends to increase as the number of trees in the forest increase, resulting in inflated type I error. The test's power does not depend on sample size either (Strobl and Zeileis 2008).

Breiman's VIMP for variable X_j is calculated through permuting X_j independently and keep the other predictors unchanged, which is analogous to sampling from the marginal distribution of X_j. As a result, variables that are not associated with the outcome but are correlated with other important variables may still have high VIMP scores because of spurious correlation (Strobl et al. 2008). To overcome this shortcoming, one may sample from the conditional distribution of X_j on all the other predictors $Z = X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$, $X_j|X_{-j}$ rather than merely permuting X_j. In this way, after the true predictors are identified, the rest would contribute little or no information to the prediction. Strobl proposed a method to sample from this conditional distribution that we call "Strobl's Conditional VIMP" (Strobl et al. 2008).

A description of the procedure is given below:

1. Before permuting X_j, calculate the out of bag prediction accuracy.

2. Conditional on Z, identify the cutpoints that split the variable in a given tree and create a grid by bissecting the variable in each cutpoint.

3. X_j is permuted within the grid; OOB prediction accuracy is calculated after the permutation.

4. VIMP of X_j within one tree is calculated by taking the difference of the OOB prediction accuracy before and after permutation. The VIMP of X_j is calculated by averaging these importance scores over all trees in the forest.

Compared with Breiman's VIMP, Strobl's conditional VIMP is less likely to assign high importance scores to variables that are highly correlated with other true predictors but not associated with outcome. The conditional variable importance has smaller variation compared with unconditional one, which makes the variable easier to be identified. However it does not solve the issue thoroughly. It still tends to assign non-zero importance to such variables that are correlated with other predictors but not the outcome (Strobl et al. 2008).

In order to solve this issue, our group proposed an alternative way of calculating conditional VIMP and a statistical test for the VIMP previously. The previous two methods provide variable importance for X_j by permuting X_j but keep other varibles unchanged, or permuting X_j within the grid of other varibles, and calculate the prediction accuracy changes before and after permutation as the variable importance. While the basic idea for the alternative conditional VIMP is the changes of prediction accuracy when X_j is replaced by the conditional distribution of X_j. When X_j is conditional important to the outcome, losing X_j or using the conditional distribution of X_j on the rest of variables would lose a lot of information and hence the prediction accuracy is low compared with that of original data. Otherwise, when X_j is not important in predicting Y, losing the information from X_j won't affect the prediction accuracy significantly. It answers the question of whether X_j brings more information in predicting Y while the other variables are present.

In this method, our group first proposed the conditional distribution of X_j on the other predictors X_{-j} through the following steps. Predict X_j by X_{-j} in a single decision tree

in random forest model. Calculate the random errors of X_j through the difference of predicted values by OOB data and the original values. Randomly permute the errors, the conditional distribution of X_j (denoted as $X_j|X_{-j}$) is derived as predicted X_j plus permuted random error. Conditional VIMP of X_j already accounts for the predictors' correlations and all the other variables. Suppose there are n subjects and p variables. Grow *ntree* trees in the forest. Perform random forest model on data to get conditional distribution of X_j in place of X_j ; caldulate the X_j's VIMP by OOB prediction accuracy difference and average them over *ntree* trees, denoted as $VIMP_j$, which is the variable importance of $X_j|X_{-j}$. In order to do the statistical test, we want to find the variable importance for X_j when it is not important to Y. A data set was generated by permuting the outcome Y so that dependent and independent variables are uncorrelated. Under the m-th tree, calculate the VIMP again. Repeat this step for B times, and let k be the k-th permutation of Y, where $k = 1, 2, \ldots, B$. The permuted VIMPs of X_j are averaged for *ntree* trees and B permutation, denoted as $\overline{VIMP}_{j,ntree,B}$. The original VIMP is shifted by subtracting $\overline{VIMP}_{j,ntree,B}$ from $VIMP_j$, denoted as $VIMP_j^*$. The column of permuted VIMP for X_j over *ntree* trees and k-th tree are denoted as $\overline{VIMP}_{j,perm}$ and $VIMP_{j,k}$ respectively. The column is centered through $\overline{VIMP}^*_{j,perm} = VIMP_{j,perm} - \frac{1}{B}\sum_{k=1}^{B} VIMP_{j,k}$. Compare $VIMP_j^*$ with $\overline{VIMP}^*_{j,perm}$ and calculate the empirical p-values for the statistical significance of VIMP score for variable j using formula $P_j = \frac{1}{p \times B}\sum_{j=1}^{p}\sum_{perm=1}^{B} I(VIMP_j^* < VIMP^*_{j,perm})$. This method does not need normality assumption. We can use small number of permutations to approximate the null distribution.

The alternative conditional VIMP method above provide a more robust way to test the variable VIMP conditional on all the other variables. It provides a way of testing if one variable are important to the outcome conditional on all the other variables. But sometimes a subset of variables might be missing simultaneously, and one may wish to know whether a group of variables bring more information to the outcome based on all the other

predictors, especially when they are highly correlated. It is possible that they as a group conditional on the rest independent variables contribute significantly to predicting outcome even though none of them individually in the subset conditional on all the others are important. They cannot be removed simmutaneously without losing prediction accuracy significantly in that case. This is very common in clinical study such as OPPERA study where a number of varibles are highly correlated and the number of variables are large. This motivated us to find out a way to test if a subset of variables are important based on the rest risk factors.

4.2 Methods

4.2.1 Conditional VIMP on a subset of variables

Similar to the alternative conditional VIMP calculation, we first proposed the conditional distribution of the variables in the subset based on all the other variables out of the subset. Suppose there are n observations and p predictors $X_1, X_2, \ldots, X_p$ with the outcome variable Y. The conditional distribution of of variables in the subset $X_{j1,\ldots,js}$ based on all the remaining predictors $X_{-j1,\ldots,-js}$ are denoted as $X_{j1}|X_{-j1,\ldots,-js}$, $\ldots$, $X_{js}|X_{-j1,\ldots,-js}$, where jk with $k = 1, \ldots, s$ is from $1, 2, \ldots, p$.

1. Fit a linear or random forest model to predict $\hat{X}_{j1}, \hat{X}_{j2}, \ldots, \hat{X}_{js}$ respectively by all other predictors $X_{-j1,\ldots,-js}$.

2. Let $\hat{x}_{i,jk}$ be the predicted value of i-th observation and jk-th covariates from OOB (out of bag) data. Then Random error $\hat{\epsilon}_{i,jk}$ was calculated by substracting the original value $x_{i,jk}$ from predicted value $\hat{x}_{i,jk}$ using OOB data when $x_{i,jk}$ is continuous. Permute the random error $\hat{\epsilon}_{i,jk}$ as $\hat{\epsilon}^*_{i,jk}$.

3. For quantative variable, let $\hat{x}^*_{i,jk} = \hat{x}_{i,jk} + \hat{\epsilon}^*_{i,jk}$, and $\hat{x}^*_{jk} = \{\hat{x}^*_{i,jk}\}$. $\hat{x}^*_{jk}$ is defined as random sample from the conditional distribution of $X_{jk}|X_{-j1,\ldots,-js}$. While the variable in

the group is qualitative, multinomial distribution was conducted to estimate the probability of each category in this variable for each subject by OOB data. Sample a value from multinomial distribution as the values estimated from OOB data.

Repeat step 2 and 3 by $k = 1, \ldots, s$. We were able to create and sample from conditional distribution of all variables in the subset. The conditional VIMP of the subset of variables was calculated from data set in the next steps adjusting for the correlation of the rest of predictors.

4.2.2 Statistical test for conditional VIMP on a subset of variables

Statistical test for the VIMP score for variable subset was also proposed. Under the null hypothesis, the conditional variable importance score for subsets $X_{j1}, ..., X_{js}$ is not significantly different from 0. When the subset is not strongly associated with the outcome, the conditional VIMP score for the subset would not change significantly after the subset are replaced by the conditional distribution of the subset. The following steps illustrate the way of calculating the conditional VIMP as well as the significant test of VIMP score for the subset predictors from the conditional distribution of the subset variables.

1. Perform random forest model and calculate conditional VIMP for the subset in a tree. First of all, calculate the OOB error by original data set through random forest. Then create new data set by replacing $X_{j1}, X_{j2}, \ldots, X_{js}$ by their conditional distributiion $\hat{X}_{j1}$, $\hat{X}_{j2}$, $\ldots$, $\hat{X}_{js}$ and keep the others unchanged. Put the new data set down the previous tree and calculate the prediction error (which is the mean square error by the difference of predicted outcome) from OOB data points in new data set and outcome in OOB data points from original data set. Then conditional VIMP of the subset in this tree was calculated by using OOB mean square error minus the original OOB error.

2. For some reason the above conditional VIMP is not close to 0 even when the subset are not associated with dependent variable. To solve this problem, we calculated the conditional VIMP when Y is independent of any Xs, and compare the conditional VIMPs. We created a data set with Y independent of all predictors by permuting dependent variable Y. B data sets were created by permuting Y for B times. In each permuted data set, similar strategy as previous step was utilized to calculate conditional VIMP in this tree. Variable X_{jk} in the l-th permutation is denoted as $X_{jk,l}$, where $l = 1, \ldots, B$. Define the outcome after the l-th permutation as Y_l, and conditional distribution of X_{jk} as $\hat{X}^*_{jk,l}$. A random forest model with one tree were perfomed on permuted data set l with all Xs as independent variables and Y_l as dependent variable. OOB error was calculated in this model. Then replace $X_{j1,l}, \ldots, X_{js,l}$ with their conditional distribution $\hat{X}^*_{j1,l}, \ldots, \hat{X}^*_{js,l}$ in data l, and put the data down the same tree in random forest to generate the predicted outcome. The OOB errors difference is the conditional VIMP score for group variables $X_{j1}, X_{j2}, \ldots, X_{js}$ in data l, denoted as $VIMP_{j,l}$. This process were repeated for B times, so that the VIMP scores of the subset for B permutation times were calculated.

3. Repeat steps 1 and 2 for *ntree* times to create *ntree* trees. The data we got contains *ntree* conditional VIMPs of the same variable subset, and *ntree* conditional VIMPs of B permuted data sets.

4. Average the conditional VIMPs over all trees and we got conditional VIMP of subset variables and B conditional VIMPs of the subset when Y is independent of all the varisbles. Let $\overline{VIMP}_j$ be the average of VIMP score for variable sets $X_{j1}, \ldots, X_{js}$ over *ntree* trees before permutation, and $\overline{VIMP}_{j,l}$ be the average of VIMP score in l-th permutation over *ntree* trees after permutation, where $l = 1, 2, \ldots, B$. $\overline{VIMP}_{j,l}$ is also averaged over B permutations as $\overline{VIMP}_{j,perm}$, where $\overline{VIMP}_{j,perm} = \frac{1}{B}\sum_{l=1}^{B} \overline{VIMP}_{j,l}$. This value denotes the noise from the data set when outcome and predictors are independent. The

shifted VIMP score ($VIMP_j$) for variable subset $X_{j1}, \ldots, X_{js}$, meaning the pure conditional VIMP from variable subset, is calculated by subtracting $\overline{VIMP}_{j,perm}$ from $\overline{VIMP}_j$, which is $VIMP_j = \overline{VIMP}_j - \overline{VIMP}_{j,perm}$. $\overline{VIMP}_{j,l}$ was centered as $VIMP^*_{j,l}$ through $VIMP^*_{j,l} = \overline{VIMP}_{j,l} - \overline{VIMP}_{j,perm}$ for $l = 1, \ldots, B$.

5. Under the null hypothesis, the subset $X_{j1}, \ldots, X_{js}$ do not have VIMP significantly different from 0 conditional on all the other predictors. The shifted VIMP score $VIMP_j$ is similar to centered VIMP score after permutation $VIMP^*_j$. P-value were obtained by comparing shifted VIMP and centered VIMP, using formula

$$P_{j1,\ldots,js} = \frac{1}{B}\sum_{l=1}^{B} I(VIMP_j < VIMP^*_{j,l}) \tag{4.16}$$

The null hypothesis is rejected when $VIMP_j$ is greater than majority of the $VIMP^*_{j,l}$s, because prediction accuracy decrease more when subset variable are associated with outcome. Conditional VIMP p-value of selected variable subset was calculated through above steps. This method can solve the problem when we want to know the if some variables can bring more information when the others are present, where these questions are pretty common in OPPERA study.

The above methods consider the outcome with quantitative traits. The following simulations include different scenarios, such as categorical varibles and outcome, where multinomial distribution was utilizef to predict categorical variables .

4.3 Simulation Studies

In this section, simulation examples were provided for the VIMP of subset variables. The significant test was also provided by calculating empirical p-values based on the method above.

4.3.1 Simulation 1, simulation of quantitative outcome and predictors

In the first simulation, both outcome and predictors were quantitative. We simulated a set of variables with 12 predictors $X_1, \ldots, X_{12}$ and 1000 subjects following multivariate normal distribution with mean 0 and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, denoted as $X \sim N(0, \Sigma)$, where $i, j = 1, 2, \ldots, 12$. The first four variables X_1, X_2, X_3, X_4 were highly correlated with covariance 0.9, but independent of other variables. The rest eight variables were mutually independent. In other words, $\{\sigma_{i,i}\} = 1$ for $i = 1, 2, \ldots, 12$; $\{\sigma_{i,j}\} = 0.9$ for $i, j = 1, 2, 3, 4$, where $i \neq j$; $\{\sigma_{i,j}\} = 0$ for all the other scenarios. The coefficients for all 12 variables were $5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 0, 0$ respectively and the outcome Y was simulated through the following formula with $\varepsilon_i \sim N(0, 0.5)$.

$$y_i = 5x_{i,1} + 5x_{i,2} + 2x_{i,3} - 5x_{i,5} - 5x_{i,6} - 2x_{i,7} + \varepsilon_i \tag{4.17}$$

In this model, $X_1, \ldots, X_4$ were highly correlated, whereas the rest eight variables were independent with each other. Among the 12 variables, only 6 were associated with the outcome, which are X_1, X_2, X_3, X_5, X_6. X_7. And X_4, X_8 to X_{12} were independent of Y. We subset variables in a group to test if they brought more information to the prediction of Y when the others were present. Method in the previous section was utilized for the test. In the random forest, 200 trees were built; the outcome Y were permuted for $B = 25$ times.

The results were listed in Table 4.2. It summarized the test statistics with p-values of conditional VIMP for the subgroups under the null hypothesis that the VIMP score for the subset variables was 0 conditional on the other variables. Variables X_1, X_2, X_3 were grouped to test if they were important to Y conditional on all the other variables. From model above, we noticed that their effect size were large compared with other predictors, and they are independent of all the other variables which are correlated with Y. They

should be important to Y. We created data with conditional distribution of three variables, and went through random forest model for the conditional VIMP with statistical test. The conditional p-value was 0 as expected, suggesting significance on 0.05 level. We rejected the null hypothesis of non-imporatance of group X_1, X_2, X_3, and they brought more information to the prediction of outcome when the other variables were present. By the same token, variables subset X_5, X_6, X_7 conditional on all other variables were significant to variable Y as well. They were independent of other variables and contribute to predicting Y. The group X_{10}, X_{11}, X_{12} had conditional p-value 0.68 which was much greater than 0.05 threshold. In the model, the coefficients for all three variables equal to 0, meaning they were not associated with Y. Hense their conditional distribution were indepent of Y and could not provide more information when the other variables were present. In the group of X_2, X_4 and X_8, the first two variables were highly correlated with each other and also associated with X_1, X_3; X_4 and X_8 were independent of Y, while X_2 was highly associated with Y. The conditional p-value was marginally significant. The similar group contains variables X_6, X_8, X_9, X_{10}, where only X_6 was associated with Y. The conditional VIMP p-value was 0, suggests significance in the test. The reason for the difference of p-values in both group is that X_2 was highly correlated with X_1 and X_3, which can be surrogates for X_2, when X_2 was missing, both of them provided some information which X_2 had. Since the effect size of X_2 was large, the p-value was marginally significant. However, X_6 was independent of all the other variables. The information from X_6 could not be provided from the other variables and hence it would lose prediction accuracy when X_6 is absent. Similarly, X_3, X_4 were conditionally not significant because they were highly correlated with X_1 and X_2. The coefficient of X_3 was not as large as the other 4 variables (X_1, X_2, X_5, X_6), meaning it did not have large effect on the outcome. X_3, X_4 could be surrogated by the other variables and do not affect prediction accuracy much. The subset X_7, X_8, X_9, X_{10} was similar to the previous setting but X_7 has small effect size to Y and

all the others are independent on Y. The subset did not show significance to the outcome.

In order to make the simulation more comprehensive, we modify the above simulation by adding interaction terms, including quanlitative predictor and binary outcome respectively.

4.3.2 Simulation 2: simulation of quantitative outcome, predictors and adding interaction term

In the second simulated model, an interaction term were added based on the first model. Similar as model 1,12 predictors were first simulated multivariate normally distributed with mean 0 and covariance matrix Σ, and six variables were associated with Y. Variables X_1, X_2, X_3, X_4 were highly correlated with each other with correlation coefficient 0.9 but independent of the other variables. All the other variables were independent of each other. The simulation of the 12 predictors were exactly the same as model 1 above. An interaction term was added so that X_{10} and X_{11} were correlated with Y only when X_{11} is greater than 0. The outcome variable was simulated by the following model.

$$y_i = 5x_{i,1} + 5x_{i,2} + 2x_{i,3} - 5x_{i,5} - 5x_{i,6} - 2x_{i,7} + \beta x_{i,10}x_{i,11}I(x_{i,11} > 0) + \varepsilon_i \qquad (4.18)$$

, where $I(x_{i,11} > 0)$ is the indicator function suggesting if $x_{i,11}$ is greater than 0 or not; when it is greater than 0 it has value 1, otherwise it has value 0.

By the simulation describe above, the data set had 1000 subjects with variables y and $x_{i,1}$ to $x_{i,12}$. The number of trees *ntree* was set to 200 and the number of permutation was 25 just the same as simulation 1. We modified β multiple times to create different value of Y and checked the conditional VIMP p-value of group X_{10}, X_{11}. The results were shown

in Table 4.3. When β was set to 4, the group conditional VIMP p-value is 0.36, suggesting not significant under 0.05 threshold. By increasing the coefficient value for the interaction term, conditional VIMP p-value kept decreasing. When β was 5, the group p-value was 0.08, indicating marginal significance. While we increased β to 6 and 20, the p-values was 0, suggesting the group variables are important to y, and the interaction of both variables played and important role in prediction Y. Consider them as a group would increase the prediction accuracy. Other combinations were tested additionally with β set to 6. Group X_1 and X_{11} showed conditional VIMP score significantly different from 0 in the test, whereas group X_{12} and X_{11} were unimportant in predicting Y. This is because X_1 was important in simulating Y, while X_{12} was independent of Y. X_{11} affected the outcome to some extent. X_{11} was independent of both variables and it did not have joint effect on outcome. Similarly subset X_3, X_{11} did not provide more information to the prediction of Y in addtion to the rest of the variables, since X_3 is highly correlated with X_1, X_2, which already provided some information of X_3, and the contribution of this group was not much. X_7 was independent of other variables. When it combined with X_{11}, the p-value suggested marginally significant in predicting Y.

4.3.3 Simulation 3, simulation with binary outcome and quantitative variables.

In simulation model 3, the outcome was binary in the model to mimic the outcome of chronic TMD in OPPERA study. The simulation of predictors were exactly the same as those in model 1, except that the outcome was binary instead of continuous. 12 predictors were simulated following multivariate normal distribution with mean 0 for all predicctors and covariance matrix Σ which was exactly the same as model 1. The first 4 variables were highly correlated with each other and independent with all the other predictors. All the other eight predictors were independent with each other. Simulate the coefficient for

all Xs the same way as these in simulation 1. In order to simulate the binary outcome Y, a linear combination of Xs, which was denoted as Z, was simulated the way in the formula shown below,

$$z_i = 5x_{i,1} + 5x_{i,2} + 2x_{i,3} - 5x_{i,5} - 5x_{i,6} - 2x_{i,7} + \varepsilon_i \tag{4.19}$$

, where i was from 1 to 1000, representing the subject ID. The predictor matrix was 1000$\times$ 12 in dimension. Let Y be the outcome of interest, following bernoulli distribution with parameter $p = 0.5$. The outcome Y followed logistic regression:

$$p(y_i = 1|x_{i,1}, \ldots, x_{i,12}) = \frac{e^{z_i}}{1 + e^{z_i}} \tag{4.20}$$

$y_i = 1$ when $p(y_i = 1|x_{i,1}, \ldots, x_{i,12}) \geq 0.5$; $y_i = 0$ when $p(y_i = 1|x_{i,1}, \ldots, x_{i,12}) < 0.5$. The number of trees *ntree* was set to 200, and the number of permutation was set to 25. The same grouping strategies were used as those in simulation 1, and we got result in Table 4.4. Groups X_1, X_2, X_3 , and X_3, X_4 were highly correlated with each other. X_1, X_2, X_3 were associated with Y and the group p-value was 0. Although X_4 was not correlated with Y, when grouped it with other imporatant variables such as X_3, the p-value still suggested importance. Similarly, when we grouped variables which were important to Y and those not important to Y together, such as the group X_2, X_4, X_8, group X_7, X_8, X_9, X_{10}, and group X_6, X_8, X_9, X_{10}, the p-values were important as well. While we group those not correlated with y such as group X_{10}, X_{11}, X_{12} the p-value was 0.12, which suggested not significant at 0.05 level.

4.3.4 Simulation 4, simulation with binary outcome and quantitative and qualitative variables.

In addition to the binary outcome, we also tried the simulation with categorical predictors based on simulation 3. A simulation of 12 variables and 1000 subjects were conducted in the following way. $x1$ to $x11$ were quantitative variables simulated the same way as those in simulation 3. Suppose x_{12} be a binary variable with values A and B. It followed bernoulli distribution of $p(A) = 0.3$. Assume $x^*_{12} = 1$ when $x_{12} = A$, and $x^*_{12} = 0$ when $x_{12} = B$, so z_i were defined as follows,

$$\begin{aligned} z_i = {} & \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6} \\ & + \beta_7 x_{i,7} + \beta_8 x_{i,8} + \beta_9 x_{i,9} + \beta_{10} x_{i,10} + \beta_{11} x_{i,11} + \beta_{12} x^*_{i,12} + \varepsilon_i \end{aligned}$$

, where i is from 1 to 1000, representing the subject ID. The binary outcome y was simulated the same way as that in simulation 3. $\beta_1, \ldots, \beta_{12}$ were simulated to be 5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 0, 0. The number of trees *ntree* was set to 200, and the number of permutation was set to 25. The subsets and corresponding p-values were shown in Table 4.5.

$$p(y_i = 1 | x_{i,1}, \ldots, x_{i,12}) = \frac{e^{z_i}}{1 + e^{z_i}} \tag{4.21}$$

$y_i = 1$ when $p(y_i = 1 | x_{i,1}, \ldots, x_{i,12}) \geq 0.5$; $y_i = 0$ when $p(y_i = 1 | x_{i,1}, \ldots, x_{i,12}) < 0.5$. The number of trees *ntree* was set to 200, and the number of permutation was set to 25. The same grouping strategies were used as those in simulation 1 as is shown in Table 4.5. The results were very similar to those in model 3. In the tested subsets listed in the table, when the group variables contained those correlated with y, such as group X_1, X_2, X_3, X_2, X_4, X_{12}, the group conditional VIMP p-values were 0s. The p-value was greater than 0.05 when we grouped X_{10}, X_{11}, X_{12} together, where none of them were correlated with the outcome.

4.4 Discussion

We focused on random forest variable selection in this chapter and provided a method for statistical testing in variable importance of a subset based on the conditional distribution of the subset on the other variables. We proposed four simulations to do the statistical tests of group variable VIMP under different scenarios, including quantitative outcome and predictors, adding interaction term, binary outcome and the continuous predictors, as well as binary outcome and mixed predictors with both continuous and categorical variables.

Based on the simulation study, when more variables in the group are highly correlated with the outcome, the statistical tests for the groups are more likely to be significant. This suggested that these varaibles are important in prediction of the outcome even when the other predictors are present. Removing them from the variable set would decrease the prediction accuracy significantly. In the simulations above, X_1, X_2, X_5, X_6 are highly correlated with the outcome, group conditional VIMP showed significance when the subsets include at least one of these variables. On the contrary, when none of the variables are associated with the outcome of interest, the statistical tests show non-significance for the group such as subset X_{10}, X_{11}, X_{12}. When some of the group variables are highly correlated with the other strong predictors, even though the we include the correlated variables with outcome in the group, the statistical test might show insignificance. It suggests that based on all the other variables, these variables cannot bring much information. The rest of variables could be the substitute to the variables in the subset. We may still get good estimate of outcome of interest without these variables. In the tests above, subset X_3, X_4 in simulations 1 and 2 showed unimportance based on the other predictors because they are highly correlated with X_1, X_2 and they do not contribute too much to the prediction of Y. In addition, the effect size of X_3 is small and X_4 is not correlated with Y.

This method solve the issue of the likelihood to assign high VIMP score to variables associated with true strong predictors but not correlated with the outcome, which existed in other methods such as Breiman's permutation VIMP and Strobl's conditonal VIMP. Statistical tests were provided to see the conditional VIMP adjusting for the other variables. It is very useful in clinical trials with many correlated variables such as OPPERA study, where hundreds of highly associated predictors were collected, and some predictors might be interchangeable (Bair et al. 2013b). Previous methods utilized in OPPERA study illustrated the univariate analysis of potential risk factors and TMD, or just the VIMP scores for risk factors. This method provides a new way to determine which groups of variable are superfluous conditional on the present variables and helps us to answer questions in OPPERA study such as whether autonomic variables are important in predicting chronic TMD conditional on all the other predictors including psychosocial, demographic and pain sensitivity variables; whether Pain Catastrophizing Scale variables have high VIMP score in predicting chronic TMD adjusting for the other variables. It provides more comprehensive information of the risk factors to both first and chronic TMD in addition to the previous study.

As we know, increasing the number of trees in the forest and the number of permutation in the calculation are useful method to decrease the variance in computation, and it will increase the computational burden. The number of trees and permutation times depends on the complexity of the data we have, such as the number of variables and subjects in the data, the correlation of predictors and etc. Choosing the balance of the two would be helpful for the test in real data. How to choose both values with respect to the complexity of the data would be the future work for this project.

Table 4.2: Simulation Results of Conditional VIMP in Variable Subset for *Simulation 1*,

Variables in The Subset Group	p-value of Conditional VIMP
X_1, X_2, X_3	0
X_5, X_6, X_7	0
X_2, X_4, X_8	0.04
X_3, X_4	0.56
X_{10}, X_{11}, X_{12}	0.68
X_7, X_8, X_9, X_{10}	0.2
X_6, X_8, X_9, X_{10}	0

Variables in The Subset Group is the set up for variable subset based on the model simulated; p-value of Conditional VIMP is the conditional VIMP p-value on all the other variables.

Table 4.3: Simulation Results of Conditional VIMP in Variable Subset for *Simulation 2*,

β	Variables in The Subset Group	p-value of Conditional VIMP
4	X_{10}, X_{11}	0.36
5	X_{10}, X_{11}	0.08
6	X_{10}, X_{11}	0
20	X_{10}, X_{11}	0
6	X_1, X_{11}	0
6	X_7, X_{11}	0.04
6	X_3, X_{11}	0.28
6	X_{12}, X_{11}	0.36

Variables in The Subset Group is the set up for variable subset based on simulated model 2; p-value of Conditional VIMP is the p-value of conditional VIMP of the subset based on all the other variables.

Table 4.4: Simulation Results of Conditional VIMP in Variable Subset for *Simulation 3*,

Variables in The Subset Group	p-value of Conditional VIMP
X_1, X_2, X_3	0
X_5, X_6, X_7	0
X_2, X_4, X_8	0
X_3, X_4	0
X_{10}, X_{11}, X_{12}	0.12
X_7, X_8, X_9, X_{10}	0
X_6, X_8, X_9, X_{10}	0

Variables in The Subset Group is the set up for variable subset based on the model simulated; p-value of Conditional VIMP is the p-value of conditional VIMP of the subset based on all the other variables.

Table 4.5: Simulation Results of Conditional VIMP in Variable Subset for *Simulation 4*,

Variables in The Subset Group	p-value of Conditional VIMP
X_1, X_2, X_3	0
X_5, X_6, X_7	0
X_2, X_4, X_{12}	0
X_3, X_4	0
X_{10}, X_{11}, X_{12}	0.72
$X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$	0
X_6, X_8, X_9, X_{12}	0

Variables in The Subset Group is the set up for variable subset based on the model simulated; p-value of Conditional VIMP is the p-value of conditional VIMP of the subset based on all the other variables.

CHAPTER 5: CONDITIONAL VIMP AND STATISTICAL TEST IN OPPERA STUDY

5.1 Introduction

Chronic pain conditions refer to pain lasting for months or longer. More than 100 million people were affected in the U.S. each year. $635 billion dollars are cost for the disease annually and it is a large burden to population (Simon 2012). The causes of chronic pain may rise from illness or injuries. Chronic pain conditions are usually treated anatomically and hence they are difficult to deal with. The examples are low back pain and Temporomandibular disorder (TMD). Even though they occur in different body parts, they share a lot of common risk factors (Diatchenko et al. 2006b). Etiology are considered in chronic pain disease and are helpful in the management and prevention in the disease. TMD is one of them. It refers to a group of painful conditions caused by the dysfunction in the jaw and related muscles (Schiffman et al. 2014). The prevalence of TMD in the U.S. is 5%, with higher prevalence among women than men (Isong et al. 2008). The average pain intensity of TMD is 4.3 out of 0 to 10 scale (Von Korff et al. 1988). Anatomic-based diagnoses are helpful in studying the disease. Other risk factors are also studied by researchers.

There are multiple risk factors in both chronic and first-onset, TMD such as clinical risk factors, sociodemographic risk factors, psychological and pain sensitivity risk factors, and genetic risk factors. They are highly correlated and the cost for the risk factors are high in both time and money. In previous study, the univariate association between baseline risk factors and the first-onset TMD incidence were conducted to illustrate the putative factors and TMD development (Slade et al. 2013), (Fillingim et al. 2013, Greenspan

et al. 2013, Ohrbach et al. 2013, Sanders et al. 2013). These are common methods in epidemiology study in finding out the causal relationship of individual risk factors and the effect.

The Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study is a prospective cohort study to identify risk factors and the etiology of TMD. The goal of the study is to find out the putative risk factors that increase the risk of TMD. The motivation of the study is a heuristic model. Basically psychological distress and the pain amplifier contribute to the risk of first-onset TMD (Maixner et al. 2011a). Participants with or without TMD aged from 18 to 44 were recruited from four U.S. study sites: Baltimore, MD; Chapel Hill, NC; Bufflo, NJ and Gainesville, FL between 2006 and 2008 through emails, flyers, advertisements, and word of mouths.

Even though the univariate and ANOVA models in previous studies casted light on the putative risk factors in TMD, the limitation cannot be ignored. The number of risk factors are large and they are highly correlated. For example, Bair et al. (2013b) discovered that different somatic awareness and autonomic variables were selected as important predictors in TMD by random forest model and lasso model; and the count of palpation sites with pain in the right masseter were selected by random forest model with high variable importance score (VIS), whereas the number of palpation sites with pain in the right and left temporalis were selected as important variables in lasso model. One may wish to find out the most vital risk factors in predicting TMD. Or people are interested in the association of a risk factor and the risk of TMD after controlling for other factors. Lasso regression and random forest modeling were used in the analysis of multiple correlated predictors in TMD. Lasso regressions penalized models with many variables to avoid overfitting. It has smaller variance than the conventional least squares model. Random forest models are based on mutliple decision trees. The decision tree is a tree graph with their consequence. Classifications and decisions can be predicted by learning the rules in the decision trees. It

is useful in non-linear model and robust in missing values but has high variance which lead to inaccurate prediction. Random forest model overcomes the limitation of decision trees by averaging them to reduce the variance of trees fitted by subsets of data. In addition to the advantage of accuracy in prediction, random forest model is able to handle missing data and large number of correlated variables accurately. Variable importance score was utilized to identiy the most important variables by random forest model. Both models were performed to answer these questions (Bair et al. 2013b). However, the limitation of both methods are obvious. They cannot provide the statistical test for the variables people are interested in. As stated in previous sections, Breiman's VIMP by permutation in random forest inclined to assign high score to variable that are associated with true important variables and may result in spurious correlation. Strobl's method by permuting variable within a grid is better than Breiman's but does not solve the issue completely (Strobl et al. 2007). We wish to find out a method to provide the statistical test to determine which subset of variables are important in predicting TMD and which variables are not when the other variables are present.

A novel method for calculating and testing the significance of group importance scores described in Chapter 4 were used to answer these questions. For example, when the group importance score for the measurements of mechanical and thermal pain sensitivity is not significantly different from 0 in predicting both chronic and first-onset TMD adjusting for the other risk factors, it implies that these measurements do not provide additional information about the risk of TMD beyond the rest risk factors. Similar methods were used to evaluate the importance of other psychological and clinical questionnaires as well as the various autonomic measures adjusting for all the other variables.

In this study, we utilized the method in Chapeter 4 to identify the subset of predictors associated with chronic TMD or first-onset TMD conditional on the other risk factors. Statistical tests were performed to identify if the groups of variables have VIMP score 0.

We want to test whether the group of varaibles are superfluous conditional on the other existing risk factors, meaning that we want to know if we lose information or not in TMD prediction even though we don't have all the information of risk factors in either chronic TMD and the first-onset TMD. Several groups of variables were identified providing more information in predicting chronic TMD after adjusting for all the other risk factors in the data set. But none of the combination we had were important in predicting first-onset TMD conditional on the rest predictors.

5.2 Methods

The data were analyzed from two observational studies, the chronic TMD and the first-onset TMD in OPPERA study. We summarized the OPPERA study, risk factors measurements and data analysis based on conditional VIMP for group variables method.

5.2.1 OPPERA study description

The "Orofacial Pain: Prospective Evaluation and Risk Assessment" (OPPERA) study is a prospective cohort study to find out the putative risk factors that affect the development of temporomandibular disorder (TMD), including psychological and physiological risk factors, the sociodemographic risk factors, clinical risk factors, autonomic risk factors and genetic mechanism. This study was funded by National Institues of Health, National Institute of Detal and Craniofacial Research. The patients were recruited from 4 study sites in the U.S., 1) The University of Maryland at Baltimore, MA, 2) The University of Buffalo, NY, 3) The University of North Carolina at Chapel Hill, NC, and 4) The University of Florida at Gainesville, FL, with the nearby population around 651k, 292k, 49k, and 95k respectively (Slade et al. 2011).

The OPPERA study consists of 4 study designs: 1) a baseline case-control study of chronic TMD, 2) a prospective cohort study of first-onset TMD, 3) a prospective cohort

study of the course of TMD and 4) the matched case-control study of incident TMD. These observational studies were designed for the identification of putative risk factors of TMD. The present analysis of baseline case-control study in chronic TMD include 4278 subjects recruited from May 2006 to May 2013 with 3247 chronic TMD controls and 1031 cases (Slade et al. 2011).

In order to find out the etiology of the first-onset of TMD, a prospective cohort study was conducted in the 4 sites targeted at people without TMD aged from 18 to 44 years old enrolling between May 2006 and November 2008 using the method of emails, advertisement and so on. 3263 participants were followed up to 5.2 years with an average of 2.8 years to ascertian the first-onset TMD. 2737 of them provided the data via questionnaires and 260 subjects were identified first-onset TMD. The annual incidence rate was 3.5%, which was determined by the total number of participants with first-onset TMD divided by the followed up person-years (Bair et al. 2013a).

The written consent was provided in both studies. The criteria of the TMD cases were using the Research Diagnostic Criteria for TMD, which include 1) $\geq$ 5 days of orofacial pain in the preceding month, and 2) during the examination, the pain is reported in TMD cases in response to the movement of one or more following muscles or joints including masticatory muscles, temporalis, submandibular and lateral pterygoid (Slade et al. 2011).

5.2.2 Study Measurenment of Risk factors in TMD

Questionnaires regarding putative related risk factors were completed by participants at enrolling. A brief description of the measurements collected in OPPERA is given below.

Sociodemographic factors and health status risk factors

Sociodemographic factors such as age, gender, race and ethnicity were collected at screen from each study participant. Additional information were reported at the baseline

visit though questionnaires, such as lifetime residence in the United States, marital status, whether the first language spoken at home is English or not, whether he/she has health insurance, education, annual household income and self-rated socioeconomic status. (Slade et al. 2013)

Potential Psychosocial Risk Factors

Previous research indicated that people with chronic TMD had higher level of the psychosocial factors such as affective distress, somatic awareness, psychosocial stress, and pain catastrophizing than TMD-free controls (Fillingim et al. 2011). In the OPPERA study, psychosocial factors were evaluated by administering questionnaires. To illustrate affective distress, the State-Trait Anxiety Inventory (STAI) and Profile of Mood States-Bipolar (POMS-Bi) were administered. STAI accesses participants' anxiety (Spielberger 1983) and POMS-Bi evaluated the positive and negative dimensions of a person's mood (Lorr 1984). The Pennebaker Inventory of Limbic Languidness (PILL) is a summary score measuring the common physical symptons and sensations frequency on five categories. (Pennebaker et al. 1982). The Kohn Reactive Scale analyzed a participant's reactivity level (Dubreuil and Kohn 1986). Both illustrate participants' somatic awareness. To get the measurement of stress, the Perceived Stress Scale (PSS), the Life Experiences Survey (LES) and the Life Stressor List/PTSD Checklist-Civilian Version (LSL/PCL-C) were administered. They evaluated stress (Cohen et al. 1983), the impact of life changes (Sarason et al. 1978) and post-traumatic stress disorder (PTSD) symptoms, respectively (Weathers et al. 1993). The Eysenck Personality Questionnaire-Revvised (EPQ-R) values personality dimension using scales of Extraversion, Neurotiicsm, and Psychoticism (Eysenck et al. 1985). The Symptiom Checklist 90-Revised (SCL-90R) evaluated global psychological distress on 9 scales such as Depression, Anxiety, Hostility, Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Phobic Anxiety, Pranoid Ideation and Psychoticism. 3 Global Indices were

also included in SCL-90R. (Derogatis 1996). The Coping Strategies Questionnaire-Revised (CSQ-R) checked the coping mechanisms used by individuals experiencing pain with 27 items (Geisser et al. 1993). The Pain Catastrophizing Scale (PCS) measured the tendency to catastrophize in response to pain (Sullivan et al. 1995). The Pittsburgh Sleep Quality Index (PSQI) is the self-reported measurement for sleep, which accesses sleep quality from 19 items during the previous month (Buysse et al. 1989).

Potential Autonomic Factors

Autonomic profiles were evaluated during the baseline visit in five periods of time: a twenty-minute rest period, a five-minute orthostatic challege period, another ten-minute rest period to access the pain-sensitivity, a five-minute color-word stroop period, and a five-minute pain-affect stroop period. Blood pressure (including systolic and diastolic blood pressure, and mean pressure, where SBP, DBP and MAP were used to name them) and heart rate (where MeanHR was used for the mean heart rate) were measured at the first period and stroop periods (which are period 4, 5) for a number of times and average them by electrocardiogram(ECG). Heart rate (HR) and blood pressure (BP) were measured prior to the first period as the initial HR and BP, as well as the second period. Several derived autonomic measures were calculated, including Total Power (TP), Very Low Frequency (VLF), Low Frequency (LF) and High Frequency (HF) HRV, which provided information about overall autonomic activity, slow temporal process activity, sympathetic activity and parasympathetic activity. The derive autonomic variables also include SDNN(standard deviation of normal-to-normal intervals) and RMSSD(root mean square of the difference between successive N-N intervals), which were time domain to measure heart rate. The derived autonomic measures were tested duing the following period except period three (Maixner et al. 2011b).

Pain Sensitivity Factors

Pain sensitivity in TMD was evaluated using psychophysical protocols. Three sensory domains were tested using psychophysical protocols. Pressure Pain Thresholds (PPT) was tested by pressure algometer using $1cm^2$ flat-tip and on five body sites (including temporalis muscles, masseter muscles, temporomandibular joint, trapezius muscles, and lateral epicondyle). Examiner increased the pressure until patients first felt pain (Greenspan et al. 2011). Mechanical Cutaneous (Pinprick) Pain were measured through weighted probes. The measurements included pain threshold, the two largest single stimulus intensities (in the response of 256mN and 512mN) with their rating of pain intesity, and temporal sumation of pain. The ratings of the pain were collected after 15 and 30 seconds of the last stimulus (Greenspan et al. 2011). Heat pain sensitivity was accessed with thermode through threshold similar to PPT by thermal stimulator. Starting at 32 °C by contacting skin in the ventral forearm, the temperature increased by 0.5 °C /sec until the participants felt pain and push bottom to record it, which estimated heat pain tolerance. Suprathreshold heat stimuli were tested repetively following the heat pain thresholds and measurement by verbally reporting the number within the range of 0 to 100 in a series of 3 stimulus with the targeted peak temperature 46 °C, 48 °C, and 50 °C respectively. (Greenspan et al. 2011). The participants recieved the temperature at a rate of 20 °C/sec for one second in order to get the hold time 750 msec for each targeted temperature. The measurements for three peak temperatures included the first thermal pulse rating, area under the curve (AUC, which is sum of rating for the pulse in each temperature) of temporal summation, and the difference of maximum pain rating and the first pulse rating, defined as "delta". The pain ratings were also collected after 15 and 30 seconds of the final pulse in the three targeted temperatures.

General Health Status Factors

Participants' general health status were accessed by questionnaires regarding health conditions for now and the past. The related checklist included endocrine, cardiovascular, hematologic, respiratory and so on. Body mass index (BMI) was measured from weight and height using formula $BMI = Weight/Height^2$. Cigarette smoking status were taken into account and classified as nonsmokers, former smokers and current smokers. Short Form 12 Health Survey v2 (SF-12v2) form was utilized to access the general health by the weighted score of physical and mental components (Sanders et al. 2013).

Clinical Factors

Various TMD clinical risk factors were assessed through baseline questionnaires. The Comprehensive Pain and Symptom Questionnaire (CPSQ) consists of a series of questions on the frequency and severity of orofacial pain and headaches as well as other comorbid pain conditions (Ohrbach et al. 2011). The Graded Chronic Pain Scale (GCPS) measured the pain intensity and interference outside orofacial region. The Jaw Functional Limitation Scale (JFLS) evaluates limitations in jaw function with respect to vertical jaw mobility and verbal/emotional expression (Ohrbach et al. 2011).

5.2.3 Statistical Analysis

Statistical analyses were conducted in both chronic TMD and first-onset TMD data sets by conditional VIMP on group variables method (illustrated in Chapter 4) to find out any combinations of variables in the group with high VIMP score significantly different from 0 in predicting TMD when the other variables are present. The number of predictors in predicting TMD cases are large, with 117 predictors in chronic TMD (with sociodemographic, psychological, pain sensitivity and autonomic risk factors) and 200 in

first on-set TMD (with clinical and health status risk factors in addition to all the risk factors in chronic TMD). Both analysis data sets exclude the predictors with more than 150 missing values to avoid computational issues. Variables with less than 150 missing values were kept in the model with all the missing values imputed by random forest. The chronic data set were from baseline case-control study described above, which include 4278 subjects with 1031chronic TMD cases and 3247 TMD free controls. The outcome is a binary variable suggesting TMD case-control status. The first-onset analysis data set contains 200 predictors and 2737 subjects with 260 cases for first-onset TMD. The outcomes were two variables, follow-up years and incidence cases, indicating how long the subjects had been followed and whether the subject was identified with TMD or cencored respectively. The follow-up years were calculated as the period between the time of enrollment to the first events of loss to follow-up or the first-onset TMD by examiner or the census date.

In this analysis, we aimed to find out the subset of variables which are enssential in elevating the risk of TMD, or groups of variables which are not important in predicting TMD risk based on all the other factors. We applied random forest conditional VIMP for group variables method described in Chapter 4 to the OPPERA data and found out groups of variables important to TMD conditional on all the other risk factors. The grouping strategies are arbitrary. Usually highly correlated variables were subset together to see if they bring more information in addtion to the other risk factors. The R package 'randomForestSRC' was used in the calculation. Statistical tests were perfomed and p-values were produced by the model for the subset variables. The nulll hypothesis H_0 was that the group variables' conditional VIMP on all the other predictors is 0. Statistical tests were performed to test the null hypothesis and conditional VIMP p-values were provided. Basically we want to see if using the conditional distribution of variables in the group in place of original variables would greatly impact the prediction accuracy.

The idea of the method is to find out if the group variables bring more information in

predicting outcome when the other variables are present. When the group variables can not bring more information in predicting TMD while the others are present, the prediction accuracy would not decrease significantly after using conditional distribution of variables in the group. The conditional distribution of a continuous variable in a group was created by the summation of predicted value by all the variables outside the group in OOB data and samples of random errors (derived from the difference between actual value and predicted value by OOB data). The conditional distribution of categorical variable in a group is the sample from multinomial distribution with probabilities estimated from variables outside the group. If the group varaibles are important, when we replace them with the conditional distribution of the variables, the prediction accuracy would decrease a lot because the group have information that cannot be covered by the other variables. The difference of prediction accuracy would indicate the VIMP score for the group. The conditional variable importance for chronic TMD was calculated by the percentage of times of predicting TMD status not equal to the original outcome from OOB data in a tree minus predicted hazard ratio while put original data down the tree, and averaged over all trees. In first-onset TMD, the conditional VIMP score was calculated similarly except that the prediction errors were calculated by $1 - concordance$. On the other hand, when the variables in the group were important to TMD, even though we have the rest of the variables, the prediction accuracy of TMD still decreased a lot after using conditional distribution of variables in the group, because they cannot be predicted accurately from the other risk factors and hence affect the prediction of outcome. Any p-values less than 0.05 are considered significant to the first-onset TMD conditional on all the other predictors. In this model, random survival forest was grown. Cox proportional hazards model was performed by the out of bag data using follow-up time and event.

In both model, the number of trees in the forest was set to 250 and the number of permutation was 100.

5.3 Results

The total 117 risk factors in baseline case-control chronic TMD study data set include 39 psychosocial risk factors, 44 autonomic risk facors, 4 sociodemographic risk factors, and 30 pain sensitivity risk factors. Besides the risk factors in chronic TMD, clinical and health status risk factors were also included in the analysis data set of first onset TMD. The strategies of grouping variables were to subset variables under the similar risk factors since they are more likely to be highly correlated with each other.

Measure of Psychosocial risk factors

The results are shown in Table 5.6. The four dimensions of EPQ-R risk factors, indicating Extraversion (EPQ-E), Neuroticism (EPQ-N), Psychoticism(EPQ-P) and Lie scale(EPQ-L) were grouped together and our hypothesis is that based on all the other risk factors, the condition variable importance score of the group of four EPQ-R variables was 0 in predicting chronic TMD. After the test, the p-value was 0.03, which was less than 0.05 threshold, indicating the VIMP for the 4 variables as a group is significantly different from 0 in predicting chronic TMD cases when the other risk factors are present. When we only grouped EPQ-N and EPQ-P in chronic TMD, the p-value was 0.92. While both test in first-onset TMD had the conditional VIMP p-values 0.27 and 0.23 respectively, which were greater than 0.05 threshold.

The group of all subcales in SCL90-R (which are variables SCL90-R in Depression, Anxiety, Hostility, Somatization, Obessive-Compulsive, Interpersonal Sensitivity, Phobic Anxiety, Pranoid Ideation and Psychotism) and Global Indices had smaller p-values conditional on the rest variables compared with those with only Depression and Obsessive-Compulsive subscales in chronic TMD analysis. The conditional p-values were 0 and 0.27 respectively. Both p-values were greater than 0.05 in the first-onset TMD, and they were 0.13 and 0.35 respectively.

STAI includes State Axiety Inventory and the Trait Anxiety Inventory. Conditional on all the other factors, the subset was not significant in predicting both chronic and the first-onset TMD with p-values 0.36 and 0.43 respectively. POMS-Bi contains 6 subscales (Agreeable-Hostile, Elated-Depressed, Confident-Unsure, Energetic-Tired, Clearheaded-Confused, and Composed-Anxious). The group conditional p-values for both TMD studies were 0.89 and 0.12 respectively. The two studies' global indices of positive an negative effects in POMS-Bi had the group conditional p-values greater than 0.05 as well, which were 0.23 for chronic TMD and 0.32 for first-onset TMD respectively. However, the measurement of Affective Distress has 8 variables, including POMS-Bi subscales and global indices, and the p-values was 0.01 indicating significant difference from 0 in VIMP in predicting chronic TMD. And the p-value of conditional VIMP of the group was 0.07 for first-onset TMD.

Psychosocial Stress measurement contains three variables, PSS, total positive and total negative events in LES in the data set. Small conditional p-value (0.02) was generated when they were grouped together in chronic TMD data. Both Pennebaker Inventory of Limbic Languidness (PILL) and Kohn Reactivity Scale (KOHN) suggested Somatic Awareness measurement, with conditional p-value 0 for a group in chronic TMD study. While the p-values in both group cases were greater than 0.05 in the first-onset TMD study.

CSQ-R has 27 items and 6 subscales indicating pain coping strategies. They are diverting attention, catastrophizing, praying and hoping, ignoring pain sensations, reinterpreting pain sensations,, and coping self-statements. The group of CSQ-R scale had p-values 0.36 and 0.26 for both Chronic TMD and first-onset conditional VIMP respectively, reflecting insignificance. Pain Catastrophizing Scale (PCS) has 3 subscales: Rumination, Magnification, Helplessness, and their p-value was 0. Combining PCS and CSQ-R risk factors for the Coping/Catastropjizing measures, the group p-value was 0 as well. Similar as the above groups in the first-onset TMD, the conditional p-values for the two groups were

greater than 0.05, and they were 0.28 and 0.17 respectively.

Table 5.6: Conditional VIMP p-values in Psychosocial Risk Factors Subsets on Chronic and First-onset TMD

Psychosocial Variables Subset Group	Conditional VIMP (p-values)	
	Chronic TMD	First-onset TMD
EPQ-P, EPQ-E, EPQ-N,EPQ-L	0.03	0.27
EPQ-P, EPQ-N	0.92	0.23
9 subscales in SCL 90-R and Global Indices	0	0.13
SCL 90-R subscales in Depression, Obsessive-Compulsive	0.27	0.35
State Anxiety Inventory and Trait Anxiety Inventory	0.36	0.43
6 subscales in POMS-Bi and 2 Global Indices of positive and negative effect	0.01	0.07
6 subscales in POMS-Bi	0.89	0.12
2 Global Indices of positive and negative effect	0.23	0.32
PSS and 2 subscales in LES	0.02	0.19
PILL and Kohn Reactive Scale	0	0.2
6 subscales in CSQ-R	0.36	0.26
3 PCS subscales	0	0.28
6 subscales in CSQ-R and 3 PCS subscales	0	0.17

Measure of Autonomic risk factors

Group p-values with respect to measuring autonomic risk factors are shown in Table 5.7. Results in chronic TMD studdies illustrated some of the groups are associated with chronic TMD. The heart rate measurement variables include HR, SPB, DPB, MAP at initial, baseline, and stroop period, as well as 5 minutes heart rate at orthostatic period. This group of variables had conditional VIMP p-value 0. Look at the group of derived variables (which are SDNN, RMSSD, LnTP, LnVLF, LnLF, and LnHF) in each period separately, the groups in period 1, 2, 4 have p-values 0s but 0.24 in period 5. The frequency domain variables include SDNN and RMSSD at period 1, 2, 4, 5, with the conditional p-value 0. However, the time domain variables contain LnVLF, LnTP, LnLF and

LnHFat period 1, 2, 4, 5, and the group p-value was 1. All the groups listed above in first-onset TMD have p-values greater than 0.05 threshold.

Table 5.7: Conditional VIMP p-values in Autonomic Risk Factors Subsets on Chronic and First-onset TMD

Autonomic Variables Subset Group	Conditional VIMP (p-values)	
	Chronic TMD	First-onset TMD
All heart rate measurement variables	0	0.16
SDNN, RMSSD, LnTP, LnVLF, LnLF, LnHF in baseline period	0	0.18
SDNN, RMSSD, LnTP, LnVLF, LnLF, LnHF in second period	0	0.33
SDNN, RMSSD, LnTP, LnVLF, LnLF, LnHF in fourth period	0	0.38
SDNN, RMSSD, LnTP, LnVLF, LnLF, LnHF in fifth period	0.24	0.43
All frequency domain variables	0	0.23
All time domain variables	1	0.24

Measure of Pain Sensitivity Risk Factors

The group strategies and conditional VIMP p-values for the pain sensitivity risk factors in chronic and first-onset TMD are described in Table 5.8.

Five body sites were tested for pressure pain thresholds (PPT). They are related to temporalis muscle, masseter mucle, temporomandibularr joint, trapezius muscle and lateral epicondyle. Conditional on all the other risk factors, PPT predictors group had conditional VIMP p-value 0 in chronic TMD analysis. When some of PPT predictors were grouped together, such as PPT trapezius and epicondyl or tramporalis and epicondyl, the conditional VIMP p-values on all the rest risk factors, including the remaining PPT variables were still 0 in chronic TMD study.

Nine pricking pain sensitivity variables, such as pain threshold, pain intensity at single stimulus by 256 and 512-mN probes, 15 and 30 seconds after sensation and residual

nonpainful sensations at both probes were grouped and tested for chronic TMD with conditional p-value 0. Conditional VIMP p-values were 0 and 0.01 with respect to the groups of variables (including pain intensity at single stimulus, after sensation at 15 and 30 seconds, and residual nonpainful sensations) under 256-mN, and 516mN probes respectively in chronic TMD analysis.

Heat pain sensitivity in pain sensitivity risk factors was accessed through heat pain tolerance, first pulse rating, delta, AUC and aftersensation pain ratings at 15 and 30 seconds in temperature 46 °C, 48 °C, and 50 °C. The following analysis were all in chronic TMD study. The conditional VIMP p-values for all heat pain sensitivity variables were 0s. The conditional p-values in the first pulse variables group, delta variables group and AUC variables group under the series of three tempearatures were 0.24, 0.89, 0.94 respectively, which were all greater than 0.05. The conditional p-value of first pulse and AUC variables was 0.18. The first pulse and delta predictors, which indicate maximum ratings under three temperatures suggested significant conditional VIMP with conditional p-value 0. The conditional VIMP p-values were $0s$ under other grouping strategies, such as the groups of all aftersensation predictors, all the heat pain predictors at 46 °C, 48 °C, and 50 °C respectively. The subset of all the heat pain predictors suggest significant conditional VIMP.

The same groups described above in chronic TMD pain sensitivity risk factors were performed and tested in first-onset TMD study. The conditional VIMP p-values were all greater than 0.05.

5.3.1 Additional Measurements in First-onset TMD

Sociodemographic risk factors such as race, gender, age group and site ID were grouped and conditional VIMP were tested in the first-onset TMD study. The p-value was 0.15.

Table 5.8: Conditional VIMP p-values in Pain Sensitivity Risk Factors Subsets on Chronic and First-onset TMD

Pain Sensitivity Variables Subset Group	Conditional VIMP (p-values)	
	Chronic TMD	First-onset TMD
PPT in all 5 body sites	0	0.29
PPT sites related to trapezius and lateral epicondyl	0	0.31
PPT sites related to tramporalis and lateral epicondyl	0	0.25
All the 9 pricking pain sensitivity variables	0	0.81
Pricking pain sensitivity variables of 256-mN	0	0.14
Pricking pain sensitivity variables of 512-mN	0.01	0.33
All the heat pain sensitivity variables	0	0.89
First pulse variables in heat pain at the target temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0.24	0.36
AUC variables in heat pain at the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0.89	0.27
Delta variables in heat pain at the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0.94	0.33
AUC and first pulse variables in the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0.18	0.29
Delta and first pulse variables in the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0	0.16
AUC, delta and first pulse variables in the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0	0.07
Aftersensation variables in heat pain at 15 seconds in the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0	0.35
Aftersensation variables in heat pain at 30 seconds in the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0	0.42
Aftersensation variables of 15 and 30 seconds in the targeted temperature $46\,^{\circ}\text{C}, 48\,^{\circ}\text{C}, 50\,^{\circ}\text{C}$	0	0.77
First pulse, AUC, delta, aftersensation variables of 15, 30 seconds in the targeted temperature $46\,^{\circ}\text{C}$	0	0.37
First pulse, AUC, delta, aftersensation variables of 15, 30 seconds in the targeted temperature $48\,^{\circ}\text{C}$	0	0.31
First pulse, AUC, delta, aftersensation variables of 15, 30 seconds in the targeted temperature $50\,^{\circ}\text{C}$	0	0.47

p-values for health status risk factors was 0.08, including smoking history, number of medications taken in the past, history of cardiovascular conditions, hematologic conditions, neural/sensory conditions, endocrine conditions, respiratory conditions, osteoarthritis, rheumatoid arthritis, Sjogrens syndrome, obstructive sleep apnea. Social economic variables group such as lifetime US residence, first spoken language, current marital status, education, finacial situation, material standards, current health insurance were grouped and testd with conditional p-value 0.34 in first-onset TMD.

5.4 Discussion

As we know, in both studies, the number of predictors are large and they are highly correlated. Measuring the variables are time and cost consuming.The study in chapters 4 and 5 provide us a way to test if some variables can provide more information in predicting TMD when the other risk factors are present. The above findings described the conditional VIMP of targeted groups based on all the rest risk factors in both baseline chronic TMD study with TMD-free controls and TMD cases, as well as baseline data in TMD-free participants when enrolled in first-onset TMD study. These findings identified some groups of varaibles in psychosocial variables, autonomic risk factors and pain sensitivity risk factors were important conditional on all the other risk factors in predicting chronic TMD. None of groups tested in first-onset TMD was important varlables group conditional on all the rest of variables. For example, PPT on 5 body sites were important to chronic TMD conditional on all the other pain sensitivity variables, autonomic, psychosocial and sociodemographic variables, but not important to first-onset TMD based on all the clinical and general health varaibles in addition to the variables described in chronic TMD.

We grouped different psychosocial risk factors and identified the groups of factors highly correlated with the case-control status of TMD conditional on the rest fo psychosocial variables and the other variables.

Consistent with previous findings that somatic awareness such that PILL, SCL 90-R differ significantly with respect to TMD case control status (Fillingim et al. 2011), the group of SCL 90-R and Global Indices, as well as the group of PILL and KOHN had p-value $0s$ in conditional VIMP. The results suggested that the groups of variables were important to TMD case-control status even though the other psychosocial risk factors were present, and they could not be omitted in the chronic TMD case-control studies. However, both groups in the first-onset TMD studies were different from previous findings, where most of variables in the univariate associations were highly correlated with the incidence rate of the first-onset TMD. None of them were important variable groups when the other predictors are present, suggesting we still get accurate first-onset TMD predictions even when either one groups are omitted. The group of SCL 90-R Depression and Obsessive-Compulsive in both chronic and first-onset TMD studies were not important. Previous studies illustrated in univariate analysis that SCL 90-R Depression was associated with TMD, and SCL 90-R Obsessive-Compulsive associated with first-onset TMD (Fillingim et al. 2011; 2013). Our study answered the questions of whether the two variables were important to either chronic or first-onset TMD when the other predictors were present. Similarly, omitting variable groups State Anxiety Inventory and Trait Anxiety Inventory would not affect both TMD prediction even though each of the variables show highly correlation with both chronic and first-onset TMD in univariate cases in previous studies (Fillingim et al. 2011; 2013).

Previous studies suggested the highly univariate association of $EPQ - N$ and low association of $EPQ - E$ with both chronic and first-onset TMD respectivley (Fillingim et al. 2011; 2013). Our tests answer the questions if the group of $EPQ - R$ or some of $EPQ - R$ variables were important to both TMD studies. The group of four $EPQ - R$ variables suggested significant conditional VIMP in chronic TMD case-control study, indicating missing all four $EPQ - R$ variables at once would greatly affect the predicting accuracy of chronic

TMD in case-control study. While only $EPQ - N$ and $EPQ - P$ were missing, we still had $EPQ - P$, $EPQ - E$, the rest of the psychosocial predictors, sociodemographic, autonomic and pain sensitivity risk factors, where we still get good prediction accuracy of chronic TMD based on them. However, in the study of first-onset TMD, both groups suggested insignificant association with TMD and hence did not affect the prediction accuracy when either groups were omitted.

The group of six subscales in POMS-Bi or the group of 2 global Indices of positive and negative effect showed insignificant conditional VIMP in predicting both chronic and first-onset TMD, although the univariate analysis suggested the two global indices individually highly associated with both chronic and first-onset TMD (Fillingim et al. 2011; 2013). Combining the above two groups together had significant conditional VIMP in chronic TMD study. This suggested that missing them simultaneously would loose information in prediction. When one group were missing, the other group could be a substitute to them and the prediction accuracy in chronic TMD could be still high. They as a group were important in predicting chronic TMD and could not be substituted by the other variables, since the prediction accuracy decreased significantly when both of them were missing. While in the first-onset TMD study, the conditional VIMP were not important to TMD whether they grouped together or not, suggesting that they were not able to bring more information in predicting TMD when the other risk factors were present. In this case, even though we don't have these variables, we still have good estimate of first-onset TMD. Similarly the group of PSS and LES brought more information in addition to the other predictors in chronic TMD but not in first-onset TMD even though the univariate analysis of LES and PSS showed significant association with first-onset TMD (Fillingim et al. 2011; 2013).

None of the following subsets, the group of Pain Catastrophizing Scale (PCS Rumination, PCS Magnification, and PCS Helplessness), the group of Coping Strategies Questionnaire (CSQ-R) risk factors and the group of both subset variables suggested insignificance in first-onset TMD. This is consistent with univariate analysis in first-onset TMD where no individual risk factors were associate with TMD. It is not surprising that the three PCS subscales group had significant conditional VIMP in chronic TMD study but CSQ-R group did not because of the highly correlated relationship of individual PCS variables with chronic TMD but only a few of CSQ-R variables associated with chronic TMD.

Previous studies showed significant association between each of the PPT variables with chronic TMD with p-values $<$ 0.0001 by analysis of variance model (Greenspan et al. 2011), indicating that PPT variables might be important to chronic TMD. Our analysis approved it by grouping all PPT variables together or part of them together, showing highly significant conditional VIMP for these variables groups. Likewise, any of the pricking pain sensitivity variables, as well as the after sensation variables in heat pain were highly associated with case-control status in ANOVA analysis in chronic TMD (Greenspan et al. 2011). Our analysis grouped all or some of the pricking pain variables and the aftersensational variables group and they showed significance in prediction chronic TMD. However, although AUC in heat pain significantly associated with chronic TMD in previous studies (Greenspan et al. 2011), the group could not bring any information when the other risk factors were present. In the first-onset TMD, most heat pain and pricking pain sensitivity variables were not highly associated with the first-onset TMD in univariate model. It is not surprising that conditional on the rest of other predictors, they could not bring new information in predicting TMD.

The derived measurements (SDNN, RMSSD, LnTP, LnVLF, LnLF, LnHF) groups suggested significant conditional VIMP in the first four period but did not bring any new information in the fifth period in predicting chronic TMD. This is probably due to the high

correlation of the last period and the previous ones. It helps us to answer the question that if the fifth period is important in terms of the derived autonomic variables conditional on the previous periods and other risk factors. We also noticed that the time domains (including LnTP, LnVLF, LnLF, LnHF in all periods) were not able to bring more information based on the other predictors even though the previous study showed significant association with chronic TMD (Maixner et al. 2011b). The frequency domains (SDNN, RMSSD) could not be eliminated in predicting chronic TMD, since they would significantly affect the prediction accuracy. Similar as before, the groups of autonomic risk factors do not bring any information when the other variables are present and hence do not affect the prediction accuracy of TMD.

In this study, conditional variable importance random forest model were utilized in both chronic TMD and first-onset TMD studies with binary outcome and time-to-event outcome respectively. Conditional p-values were calculated for the statistical testing of variable subset. Some variable subsets were identified significantly in predicting chronic TMD. None of the subsets were important in first-onset TMD. One reason is smaller number of data set but larger number of variables in first-onset TMD data set than chronic TMD data set. Their number of variables and participants differences are 117 VS. 200 and 4278 VS. 2737 in chronic TMD and first-onset TMD data set, respectively. Each variable in chronic TMD might contribute more than the first-onset TMD. The second reason is that when we looked at the previous findings regarding each variable and TMD, more variables are found associated with chronic TMD than first-onset TMD by univariate model. When we combined them together and rule out the other predictors' affect, it is likely that more variable groups are associated with chronic TMD. Thirdly, the variables are highly correlated with each other, interactions among variables are likely in the model. The current analyses did not include interaction terms in both study. Further research might be needed in terms of interaction terms. Fourthly, the incidence cases in first-onset TMD

study is 260, which is relatively small, and the power of the testing is limited as well.

Another interesting finding is that increasing the number of trees in random forest to 250 and the number of permutation to 100, which are typically reducing variance and increace accuracy, would make the p-values for chronic TMD more stable than those for first-onset TMD data. We ran each group's analysis for 10 times using different seed and notice that the p-values in a group for chronic TMD are the same but not for first-onset TMD. The p-values are more likely to be random distributed. Each varible in first-onset TMD contribute little to TMD, when they were removed the prediction accuracy is not affected significantly, which were consistent with most of the previous studies. We probably would increase the number of trees or the number of permutations in the second data set, but it would increase the computation burden for sure.

In this chapter, we discussed the risk factors in TMD of both baseline case-control study and prospective cohort study. we studied a number of questions in the analysis of predictors as a group when the other predictors were present, and computed the p-values for the statistical analysis. We answered the questions such as if all heart rate measurements are important variables in predicting both chronic and first-onset TMD, whether AUC variables in heat pain sensitivity risk factors can be eliminated without increasing prediction error. This is a great help to predict TMD accurately when some variables are missing or only a subset of variables are present due to limited budget.

CHAPTER 6: SUMMARY AND FUTURE RESEARCH

The Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study is a prospective study designed to study the etiology and find out the risk factors contributing to the onset and chronic temporomandibular disorder (TMD). Numerous risk factors were studied and identified for TMD, such as pain sensitivity risk factors, clinical risk factors, psychological distress, and genetic factors. In this dissertation, three projects were included to study the genetic risk factors, and the other risk factors as a group by using nonparametric and machine learning methodologies.

The first topic focused on the question of genetic association with secondary outcome which was highly associated with case control status. Permutation-based IPW method were proposed to address this question. Compared with conventional IPW method which produced inflated p-values, our nonparametric method do not have inflated type I errors. The power of the test were comparable to conventional IPW method. Our method was applied to OPPERA study to find out the association of TMD pain density and candidate genetic variants, where two SNPs were identified associated with secondary outcome. However, this method only produces p-values. It is not able to calculate odds ratios or other effect size. As we know that conventional IPW method produced anticonservative results in the variance of effect size, future research is need to address the question.

In the second and third topic, we focus on the variable importance by random forest in OPPERA study. We developed a novel method by testing whether the group of risk factors' variable importance is 0 adjusting for the existing variables. We tested for a number

of groups of predictors to see that based on the variables we have, would the group variables bring more information. This method corrects the shortcomings of the likelihood of choosing correlated variables with spurious correlation, and provided a way of testing group variables without bias. The method were applied to OPPERA study to test the risk factors for chronic and first-onset TMD, where many groups of variables were identified importance to chronic TMD conditional on the rest risk factors, whereas none of the groups were found significant to first-onset TMD adjusting for the rest predictors.

REFERENCES

Archer, K. J. and Kimes, R. V. (2008), "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, 52, 2249–2260.

Austin, M. A., Beaty, T. H., and Dotson, W. D. (2013), *Genetic Epidemiology: Methods and Applications*, CABI.

Bair, E. (2013), "Identification of significant features in DNA microarray data," *Wiley Interdisciplinary Reviews: Computational Statistics*, 5, 309–325.

Bair, E., Brownstein, N. C., Ohrbach, R., Greenspan, J. D., Dubner, R., Fillingim, R. B., Maixner, W., Smith, S. B., Diatchenko, L., Gonzalez, Y., et al. (2013a), "Study protocol, sample characteristics, and loss to follow-up: The OPPERA prospective cohort study," *The Journal of Pain*, 14, T2–T19.

Bair, E., Gaynor, S., Slade, G. D., Ohrbach, R., Fillingim, R. B., Greenspan, J. D., Dubner, R., Smith, S. B., Diatchenko, L., and Maixner, W. (2016), "Identification of clusters of individuals relevant to temporomandibular disorders and other chronic pain conditions: the OPPERA study," *Pain*, 157, 1266–1278.

Bair, E., Ohrbach, R., Fillingim, R. B., Greenspan, J. D., Dubner, R., Diatchenko, L., Helgeson, E., Knott, C., Maixner, W., and Slade, G. D. (2013b), "Multivariable modeling of phenotypic risk factors for first-onset TMD: the OPPERA prospective cohort study," *The Journal of Pain*, 14, T102–T115.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Breiman, L. (2001), "Random forests," *Machine learning*, 45, 5–32.

Breiman, L. and Cutler, A. (2008), "Random forests–Classification manual. Website accessed in 1/2008," .

Bush, W. S. and Moore, J. H. (2012), "Chapter 11: Genome-wide association studies," *PLoS Comput Biol*, 8, e1002822.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., and Kupfer, D. J. (1989), "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research," *Psychiatry research*, 28, 193–213.

Cairns, B. (2010), "Pathophysiology of TMD pain–basic mechanisms and their implications

for pharmacotherapy," *Journal of oral rehabilitation*, 37, 391–410.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010), "Prioritizing GWAS results: a review of statistical methods and recommendations for their application," *The American Journal of Human Genetics*, 86, 6–22.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983), "A global measure of perceived stress," *Journal of health and social behavior*, 385–396.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007), "Random forests for classification in ecology," *Ecology*, 88, 2783–2792.

de Dieu Tapsoba, J., Kooperberg, C., Reiner, A., Wang, C.-Y., and Dai, J. Y. (2014), "Robust estimation for secondary trait association in case-control genetic studies," *American journal of epidemiology*, kwu039.

Derogatis, L. R. (1996), *SCL-90-R: Symptom Checklist-90-R: administration, scoring, and procedures manual*, NCS Pearson.

Diatchenko, L., Anderson, A. D., Slade, G. D., Fillingim, R. B., Shabalina, S. A., Higgins, T. J., Sama, S., Belfer, I., Goldman, D., Max, M. B., et al. (2006a), "Three major haplotypes of the $\beta 2$ adrenergic receptor define psychological profile, blood pressure, and the risk for development of a common musculoskeletal pain disorder," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141, 449–462.

Diatchenko, L., Nackley, A. G., Slade, G. D., Fillingim, R. B., and Maixner, W. (2006b), "Idiopathic pain disorders–pathways of vulnerability," *Pain*, 123, 226–230.

Díaz-Uriarte, R. and De Andres, S. A. (2006), "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, 7, 3.

Dubreuil, D. L. and Kohn, P. M. (1986), "Reactivity and response to pain," *Personality and Individual Differences*, 7, 907–909.

Dworkin, S. F., Huggins, K. H., LeResche, L., Von Korff, M., Howard, J., Truelove, E., and Sommers, E. (1990), "Epidemiology of signs and symptoms in temporomandibular disorders: clinical signs in cases and controls," *The Journal of the American Dental Association*, 120, 273–281.

Etoz, O. A., Ataoglu, H., and Erdal, M. E. (2008), "Association between tryptophan hydroxylase gene polymorphism and painful non-osseous temporomandibular disorders." *Saudi medical journal*, 29, 1352–1354.

Eysenck, S. B., Eysenck, H. J., and Barrett, P. (1985), "A revised version of the psychoticism scale," *Personality and individual differences*, 6, 21–29.

Fillingim, R. B., Ohrbach, R., Greenspan, J. D., Knott, C., Diatchenko, L., Dubner, R., Bair, E., Baraian, C., Mack, N., Slade, G. D., et al. (2013), "Psychological factors associated with development of TMD: the OPPERA prospective cohort study," *The Journal of Pain*, 14, T75–T90.

Fillingim, R. B., Ohrbach, R., Greenspan, J. D., Knott, C., Dubner, R., Bair, E., Baraian, C., Slade, G. D., and Maixner, W. (2011), "Potential psychosocial risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case-control study," *The Journal of Pain*, 12, T46–T60.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., et al. (2007), "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity," *Science*, 316, 889–894.

Geisser, M. E., Robinson, M. E., and Pickren, W. E. (1993), "Differences in cognitive coping strategies among pain-sensitive and pain-tolerant individuals on the cold-pressor test," *Behavior Therapy*, 23, 31–41.

Ghosh, A., Wright, F. A., and Zou, F. (2013), "Unified Analysis of Secondary Traits in Case–Control Association Studies," *Journal of the American Statistical Association*, 108, 566–576.

Goldstein, B. A., Hubbard, A. E., Cutler, A., and Barcellos, L. F. (2010), "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings," *BMC genetics*, 11, 49.

Greenspan, J. D., Slade, G. D., Bair, E., Dubner, R., Fillingim, R. B., Ohrbach, R., Knott, C., Diatchenko, L., Liu, Q., and Maixner, W. (2013), "Pain sensitivity and autonomic factors associated with development of TMD: The OPPERA prospective cohort study," *The Journal of Pain*, 14, T63–T74.

Greenspan, J. D., Slade, G. D., Bair, E., Dubner, R., Fillingim, R. B., Ohrbach, R., Knott, C., Mulkey, F., Rothwell, R., and Maixner, W. (2011), "Pain sensitivity risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case control study," *The Journal of Pain*, 12, T61–T74.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Series in Statistics, New York, NY: Springer, 2nd ed.

He, C., Kraft, P., Chen, C., Buring, J. E., Paré, G., Hankinson, S. E., Chanock, S. J., Ridker, P. M., Hunter, D. J., and Chasman, D. I. (2009), "Genome-wide association studies identify loci associated with age at menarche and age at natural menopause," *Nature genetics*, 41, 724–728.

He, J., Li, H., Edmondson, A. C., Rader, D. J., and Li, M. (2011), "A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies," *Biostatistics*, kxr025.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., et al. (2008), "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25," *Nature*, 452, 633–637.

Isong, U., Gansky, S. A., and Plesh, O. (2008), "Temporomandibular joint and muscle disorder-type pain in US adults: the National Health Interview Survey," *Journal of orofacial pain*, 22, 317.

Kammerer, C. M., Gouin, N., Samollow, P. B., VandeBerg, J. F., Hixson, J. E., Cole, S. A., MacCluer, J. W., and Atwood, L. D. (2004), "Two quantitative trait loci affect ACE activities in Mexican-Americans," *Hypertension*, 43, 466–470.

Karayiorgou, M., Sobin, C., Blundell, M. L., Galke, B. L., Malinova, L., Goldberg, P., Ott, J., and Gogos, J. A. (1999), "Family-based association studies support a sexually dimorphic effect of COMT and MAOA on genetic susceptibility to obsessive-compulsive disorder," *Biological psychiatry*, 45, 1178–1189.

Kwon, J. M. and Goate, A. M. (2000), "The candidate gene approach," *Alcohol Research and Health*, 24, 164–168.

Lawford, B. R., Young, R. M., Noble, E. P., Kann, B., Arnold, L., Rowell, J., and Ritchie, T. L. (2003), "D2 dopamine receptor gene polymorphism: paroxetine and social functioning in posttraumatic stress disorder," *European neuropsychopharmacology*, 13, 313–320.

Lewis, C. M. and Knight, J. (2012), "Introduction to genetic association studies," *Cold Spring Harbor protocols*, 2012, pdb–top068163.

Li, H., Gail, M. H., Berndt, S., and Chatterjee, N. (2010), "Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies," *Genetic epidemiology*, 34, 427–433.

Li, Q. and Yu, K. (2008), "Improved correction for population stratification in genome-wide

association studies by identifying hidden population structures," *Genetic epidemiology*, 32, 215–226.

Lin, D. and Zeng, D. (2009), "Proper analysis of secondary phenotype data in case-control association studies," *Genetic epidemiology*, 33, 256–265.

Liu, L., Zhang, D., Liu, H., and Arendt, C. (2013), "Robust methods for population stratification in genome wide association studies," *BMC bioinformatics*, 14, 132.

Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S., et al. (2008), "Common variants near MC4R are associated with fat mass, weight and risk of obesity," *Nature genetics*, 40, 768–775.

Lorr, M. (1984), *Profile of Mood States: bi-polar form (POMS-BI): manual*, Educational and Industrial Testing Service.

Maixner, W., Diatchenko, L., Dubner, R., Fillingim, R. B., Greenspan, J. D., Knott, C., Ohrbach, R., Weir, B., and Slade, G. D. (2011a), "Orofacial pain prospective evaluation and risk assessment study–the OPPERA study," *The Journal of Pain*, 12, T4–T11.

Maixner, W., Greenspan, J. D., Dubner, R., Bair, E., Mulkey, F., Miller, V., Knott, C., Slade, G. D., Ohrbach, R., Diatchenko, L., et al. (2011b), "Potential autonomic risk factors for chronic TMD: descriptive data and empirically identified domains from the OPPERA case-control study," *The Journal of Pain*, 12, T75–T91.

Monsees, G. M., Tamimi, R. M., and Kraft, P. (2009), "Genome-wide association scans for secondary traits using case-control samples," *Genetic epidemiology*, 33, 717–728.

Mutlu, N., Erdal, M., Herken, H., Oz, G., and Bayazıt, Y. (2004), "T102C polymorphism of the 5-HT2A receptor gene may be associated with temporomandibular dysfunction," *Oral diseases*, 10, 349–352.

Ohrbach, R., Bair, E., Fillingim, R. B., Gonzalez, Y., Gordon, S. M., Lim, P.-F., Ribeiro-Dasilva, M., Diatchenko, L., Dubner, R., Greenspan, J. D., et al. (2013), "Clinical orofacial characteristics associated with risk of first-onset TMD: the OPPERA prospective cohort study," *The Journal of Pain*, 14, T33–T50.

Ohrbach, R., Fillingim, R. B., Mulkey, F., Gonzalez, Y., Gordon, S., Gremillion, H., Lim, P.-F., Ribeiro-Dasilva, M., Greenspan, J. D., Knott, C., et al. (2011), "Clinical findings and pain symptoms as potential risk factors for chronic TMD: descriptive data and empirically

identified domains from the OPPERA case-control study," *The Journal of Pain*, 12, T27–T45.

Pennebaker, J. W., Gonder-Frederick, L., Stewart, H., Elfman, L., and Skelton, J. (1982), "Physical symptoms associated with blood pressure," *Psychophysiology*, 19, 201–210.

Plesh, O., Crawford, P. B., and Gansky, S. A. (2002), "Chronic pain in a biracial population of young women," *Pain*, 99, 515–523.

Plesh, O., Noonan, C., Buchwald, D. S., Goldberg, J., and Afari, N. (2012), "Temporomandibular disorder-type pain and migraine headache in women: a preliminary twin study." *Journal of orofacial pain*, 26.

Posthuma, D., de Koning, D.-J., Dolan, C., Goddard, M. E., and Visscher, P. M. (2009), "A note on permutation tests for genetic association analysis of quantitative traits when variances are heterogeneous," *Genetic epidemiology*, 33, 710–716.

Prentice, R. L. and Pyke, R. (1979), "Logistic disease incidence models and case-control studies," *Biometrika*, 66, 403–411.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006), "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, 38, 904–909.

Pritchard, J. K. and Donnelly, P. (2001), "Case–control studies of association in structured or admixed populations," *Theoretical population biology*, 60, 227–237.

Richardson, D. B., Rzehak, P., Klenk, J., and Weiland, S. K. (2007), "Analyses of case–control data for additional outcomes," *Epidemiology*, 18, 441–445.

Sanders, A. E., Slade, G. D., Bair, E., Fillingim, R. B., Knott, C., Dubner, R., Greenspan, J. D., Maixner, W., and Ohrbach, R. (2013), "General health status and incidence of first-onset temporomandibular disorder: the OPPERA prospective cohort study," *The Journal of Pain*, 14, T51–T62.

Sarason, I. G., Johnson, J. H., and Siegel, J. M. (1978), "Assessing the impact of life changes: development of the Life Experiences Survey." *Journal of consulting and clinical psychology*, 46, 932.

Schiffman, E., Ohrbach, R., Truelove, E., Look, J., Anderson, G., Goulet, J.-P., List, T., Svensson, P., et al. (2014), "Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: recommendations of the International

RDC/TMD Consortium Network and Orofacial Pain Special Interest Group," *Journal of oral & facial pain and headache*, 28, 6.

Simon, L. S. (2012), "Relieving pain in America: A blueprint for transforming prevention, care, education, and research," *Journal of Pain & Palliative Care Pharmacotherapy*, 26, 197–198.

Slade, G., Diatchenko, L., Bhalang, K., Sigurdsson, A., Fillingim, R., Belfer, I., Max, M., Goldman, D., and Maixner, W. (2007), "Influence of psychological factors on risk of temporomandibular disorders," *Journal of dental research*, 86, 1120–1125.

Slade, G. D., Bair, E., By, K., Mulkey, F., Baraian, C., Rothwell, R., Reynolds, M., Miller, V., Gonzalez, Y., Gordon, S., et al. (2011), "Study methods, recruitment, sociodemographic findings, and demographic representativeness in the OPPERA study," *The Journal of Pain*, 12, T12–T26.

Slade, G. D., Bair, E., Greenspan, J. D., Dubner, R., Fillingim, R. B., Diatchenko, L., Maixner, W., Knott, C., and Ohrbach, R. (2013), "Signs and symptoms of first-onset TMD and sociodemographic predictors of its development: The OPPERA prospective cohort study," *The Journal of Pain*, 14, T20–T32.

Smith, L. I. (2002), "A tutorial on principal components analysis," *Cornell University, USA*, 51, 52.

Smith, S. B., Maixner, D. W., Greenspan, J. D., Dubner, R., Fillingim, R. B., Ohrbach, R., Knott, C., Slade, G. D., Bair, E., Gibson, D. G., et al. (2011), "Potential genetic risk factors for chronic TMD: genetic associations from the OPPERA case control study," *The Journal of Pain*, 12, T92–T101.

Spielberger, C. D. (1983), "Manual for the State-Trait Anxiety Inventory STAI (form Y)(" self-evaluation questionnaire")," .

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional variable importance for random forests," *BMC bioinformatics*, 9, 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, 8, 1.

Strobl, C. and Zeileis, A. (2008), "Danger: High power!–exploring the statistical properties of a test for random forest variable importance," .

Sullivan, M. J., Bishop, S. R., and Pivik, J. (1995), "The pain catastrophizing scale: development and validation." *Psychological assessment*, 7, 524.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002), "Candidate-gene approaches for studying complex genetic traits: practical considerations," *Nature Reviews Genetics*, 3, 391–397.

Tian, C., Gregersen, P. K., and Seldin, M. F. (2008), "Accounting for ancestry: population substructure and genome-wide association studies," *Human molecular genetics*, 17, R143–R150.

Villeneuve, P. J. and Mao, Y. (1993), "Lifetime probability of developing lung cancer, by smoking status, Canada." *Canadian journal of public health= Revue canadienne de sante publique*, 85, 385–388.

Von Korff, M., Dworkin, S. F., Le Resche, L., and Kruger, A. (1988), "An epidemiologic comparison of pain complaints," *Pain*, 32, 173–183.

Wang, J. and Shete, S. (2011), "Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases," *Genetic epidemiology*, 35, 190–200.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., Keane, T. M., et al. (1993), "The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility," in *Annual Convention of the International Society for Traumatic Stress Studies*, International Society for Traumatic Stress Studies San Antonio.

WuÌĺst, S., Van Rossum, E. F., Federenko, I. S., Koper, J. W., Kumsta, R., and Hellhammer, D. H. (2004), "Common polymorphisms in the glucocorticoid receptor gene are associated with adrenocortical responses to psychosocial stress," *The Journal of Clinical Endocrinology & Metabolism*, 89, 565–573.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010), "Mixed linear model approach adapted for genome-wide association studies," *Nature genetics*, 42, 355–360.