Coping With Complexities in High Dimensional Data: PheWAS in EMR and Statistical Inference in fMRI Data

By

Ya-Chen Lin

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Biostatistics

August 7, 2020

Nashville, Tennessee

Approved:

Robert Johnson, Ph.D.

Hakmook Kang, Ph.D.

Yaomin Xu, Ph.D.

Todd Edwards, Ph.D.

This dissertation is dedicated to my beloved mom, Natasha.
My strongest and most solid supporter.
Mom, I made it!

ACKNOWLEDGEMENTS

Coming from biochemistry background, switching major to Biostatistics was a hard transition. Fortunately, Biostatistics department in Vanderbilt University had noticed my determination and rich knowledge in the biological science. The past five years were indeed challenging but also rewarding.

I want to say a big thank you to my main advisor, Hakmook. The road to becoming an independent researcher is hard. You have offered me tremendous help along the way while providing me with freedom to conduct research that I enjoy. You helped make this journey smoother and deepen my passion in Biostatistics. I also appreciate all your advice regarding to my future career path.I I also want to thank Dr.Yaomin Xu from Biostatistics department, who serves as a co-mentor on the Chapter 2 project for all his helpful advice and suggestions on the project.

A special thanks to my committee chair, Bob. Not only did you provide me with academic advice when I can't figure out questions related to math but also offered me with guidelines and suggestion when I felt lost during the PhD journey, whether they were related to academic decision-making, research interest or future planning. I thank Dr. Todd Edwards from Department of Medicine as you have provided me with suggestions during committee meetings.

For dissertation contents, I want to thank Dr. Jeffrey Blume and Valerie Welty for all the helpful edits for Chapter 4 on second-generation p-values. I also want to thank Dr. Warren Taylor's lab, especially Dr. Warren Talyor, Dr. Kimberly Albert. I want to also thank the Dr. Josh Denny's lab for the information regarding to PheWAS and awesome PheWAS R package.

I want to especially thank my family, Natasha (mom), Jack (dad) and Alec (brother). Being away from home country for almost 10 years, if it weren't for your support, I would not have made it here. Your endless cheers and phone calls help me through multiple frustrating nights when I miss home the most. Thank you for listening to me rambling about my dissertation progress and giving consoles when I need them the most. You guys are my best and most solid supporters!

Lastly, I want to thank my classmates/friends in Biostatistics department. We share with each other the difficulties and happiness on the road to becoming a PhD. We cheer for and console each other. I also thank my friends around the world, even though we don't see each other very often, I feel connected to all of you. A phone call or a text message always make me feel connected.

Thank you all, my advisor, family, committee members and friends. I won't be able to achieve what I have without your support. You guys are the best!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC	Anterior commissure
AR(1)	First order autoregressive model
BOLD	Blood Oxygen Level Dependent
CSF	Cerebrospinal fluid
D-SecondP	Dichotomizded version of Second-generation p-value
EDP	Emotion dot probe
EMR	Electronic medical record
FDR	False discovery rate
fMRI	functional Magnetic resonance imaging
FOV	Field of view
FSL	FMRIB Software Library
FWHM	Full width, half maximum
GLM	General linear model
GWAS	Genome-Wide Association Study
HRF	Hemodynamic response function
IAPS	International Affective Pictures System
IQR	Interquartile range
MDD	Major depression disorder
MNI	Montreal Neurological Institute
MPRAGE	Magnetization-prepared rapid gradient-echo
NIfTI	Neuroimaging Informatics Technology Initiative
NSA	Number signal averages

PC	posterior commisure
PheWAS	Phenome-Wide Association Study
RFT	Random field theory
ROI	Region of interest
SGPV/SecondP	Second-generation p-value
SNP	single nucleotide polymorphism
SNR	Signal-to-noise ratio
SPM	Statistical parametric mapping
TE	Echo time
TFESENSE	Turbo field echo sensitivity encoding
TR	Repetition time
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

This dissertation aims to thoroughly explore some complexities one can face while handling large dimensional data and provide some ready-to-use alternatives to address critical issues. We specifically focus on electronic medical record (EMR) data in Chapter 2 and task-induced functional magnetic resonance imaging (fMRI) data in Chapters 3 - 4. In this introduction Chapter, we provide some general backgrounds that are directly related to the data we explore and conclude with an outline of the work proposed in Chapters 2 - 4.

1.1 Phenome-Wide Association Studies (PheWAS) in EMR Data

Similar to Genome-wide association studies (GWAS), Phenome-wide association studies scan through the phenotypes in the database with specific genotype of interest. The goal for PheWAS is to explore disease comorbidities related to the genotype of interest and can benefit the downstream analyses (i.e., drug repurpose or targeted treatments). Phecodes, billing code-derived disease case-control status, are usually used as binary outcome variables in PheWAS and logistic regression has been the standard choice of analysis method. Since the clinical diagnoses in EMR are often inaccurate with errors, which can lead to biases in the odds ratio estimates, much effort has been put to accurately define the cases and controls to ensure an accurate analysis.

1.1.1 The Need for Exclusion Criteria Lists

Denny et al. (2016) realized this limitation and tried to come up with an automatic process to correctly classify the controls. Specifically, in order to correctly classifying controls in the population, an exclusion criteria list for each Phecode was manually compiled by a group of physician to obtain unbiased odds ratios. The proposed PheWAS diagnosis diagram is as follow. If a subject has two or more ICD 9 codes on different days that map to the Phecode of interest, the subject is categorized as a case. If a subject has only 1 ICD 9 that map to the Phecode of interest, the subject is excluded from the analysis. If a subject does not have any ICD 9 code that maps to the Phecode of interest nor any ICD 9 code on the exclusion criteria list, the subject is then categorized as a true control. Otherwise, the subject is excluded for the analysis.

However, even with only 1800 Phecodes, it took around two years or more to fully compile the exclusion criteria lists for all Phecodes. In addition, the comorbidities of the diseases are still largely unknown. Therefore, the accuracy of the list cannot be guaranteed without extensive data curation process. The costly curation process limits the efficiency of large-scale analyses that take full advantage of all structured phenotypic information available in EMR.

1.1.2 Study Designs in PheWAS with EMR Data

In Chapter 2, we focus on the study designs in EMR data. In general, there are two well-known study designs. The first design is the case-control design where the investigators collect data based on the outcome of interest. The exposure status of each subject is also recorded. The second design is the cohort design where data are collected based on the exposure of interest. Each subject is followed up to examine the case status, prospectively or retrospectively. In PheWAS, the genotype of interest can be viewed as the exposure and the various phenotypes can be viewed as outcomes. Therefore, instead of case-control studies, the study design in PheWAS resembles a retrospective cohort design more. We later show that the desired nature of a different estimator, valid in cohort studies, allows us to bypass the need for exclusion criteria lists. PheWAS can be efficiently extended to a larger-scale, phenome construction agnostic analysis of phenotypes, which use ICD 9/10 codes, preserve much more disease-related clinical information than Phecodes.

1.2 Task-induced fMRI Data Analysis

Functional magnetic resonance imaging (fMRI) is a tool that measures brain activities by detecting dynamic changes associated with blood flow. As the neuronal activities of a brain region increase, the oxygen consumption and the blood flow increase. The blood-oxygen-level-dependent (BOLD) signal then serves as an indirect measurement for neuronal activities in fMRI data. In this dissertation, we focus on task-induced fMRI data which measure the contrasts between stimuli of interest. With task-induced fMRI analysis, the investigators can learn about the regions of the brain associated with specific task of interest.

1.2.1 Pipeline to preprocess the data

To obtain task-induced fMRI brain images, multiple subjects are usually recruited. These subjects are asked to perform a specific task (i.e., finger tapping or reacting to different images). For each subject, the brain is scanned by the MRI machine for BOLD signals at various time points. At the end, we can collect the time series of BOLD signals for each subject. Before we can begin the analysis, several preprocessing steps must be performed. These preprocessing steps are usually standardized but can vary between different experiments. Some major steps include slice-time correction, motion correction, co-registration, normalization, high and low-pass temporal filtering. These steps ensure the processed data are free of any contamination that is not directly related to neural activity. Lastly, spatial smoothing is often applied before the analysis. Due to the noisy nature of the data, spatial smoothing is an important preprocessing step to increase the signal-to-noise ratio (SNR). After a series of preprocessing steps, the data are ready for analysis. Due to the large number of comparisons, multiple correction methods are usually used to obtain appropriate statistical inference in fMRI data analysis. Common multiple correction methods include controlling for family-wise error rates (random field theory, RFT) and false discovery rates (FDR). Controlling for family-wise error rates provide strong control over the number of false positives, but tends to be conservative, leading to low power. Controlling for FDR is more liberal and preferable method nowadays to better balance the false positives and power. Both methods rely on p-values.

1.2.2 Second-generation p-values

There are several drawbacks of p-value including interpretation issues and separation with clinical significance. At first, p-value was created to measure the probability of obtaining the result or more extreme based on current data assuming null hypothesis is true. After being incorporated as a measurement for hypothesis testing, p-value is sometimes misinterpreted as Type I error rate or false discovery rate, leading to confusion of the definition of p-value. More and more discussion has been surrounding the difference between statistical and clinical significance. In traditional hypothesis testing, null hypothesis is set at a point (i.e., 0). When the result rejects the null hypothesis, there might be points in the alternative hypothesis that are not clinically different from the null. We call the result statistically significant but not necessarily clinically meaningful. In order to address these drawbacks of p-value, Blume et al. (2018) introduced second-generation p-values (SGPV). By expanding the null hypothesis from a point to an interval, points that are not clinically meaningful are incorporated as an interval null for hypothesis testing. SGPV is simply the proportion of observed data that support the null hypothesis. With convenient interpretation and the incorporation of interval null, SGPV overcomes the long-standing drawbacks of p-values. In our dissertation, we introduce SGPV as a new inference tool in task-induced fMRI analysis to better balance between Type I and II error rates.

1.3 Dissertation Focus

Due to the large dimensionalities of the data we focus on, regular statistical methods might not work well, forcing the researchers to develop new techniques to account for it. However, the additional noisy nature of the biomedical large dimensional data makes statistical analyses even more complex.

1.3.1 General problems

In PheWAS analysis, the prevalence of the outcomes varies largely, leading to difficulties in convergence of more sophisticated new methods. These newly-developed methods are usually computationally burdensome and the convergence of the models cannot be guaranteed. Therefore, efficient estimation methods are critically in need. In task-induced fMRI data analysis, multiple preprocessing steps are required to reduce and correct for noise. These steps are usually necessary and standardized in researcher's analysis pipelines. However, these steps, if not carefully considered, could be influential on statistical inference. The frequentist inference approaches in task-induced fMRI data analysis tend to be either far too conservative or they fail to correct for multiple comparisons. Novel methods have been proposed. These methods are mostly computational burdensome and focus solely on ensuring the correct nominal family-wise Type I error rates. However, more and more literature (Slotnick (2017)) are now claiming that the ignorance of Type II error rates could severely hinder the scientific discoveries. A good method should balance both Type I and II error rates simultaneously.

1.3.2 A glance of contributions

- We propose to view the study design in PheWAS as retrospective cohort design and provide a different estimator which allows PheWAS to be extended to analyzing larger-scale phenotypes efficiently (Chapter 2).

- We thoroughly explore the influence of the choice of spatial smoothing along with experimental factors on Type I and II error rates. We also extend the spatial smoothing to maximum likelihood estimates (Chapter 3).
- We introduce a novel inferential method, second-generation p-values (SGPV) that improves upon the p-values inferential framework and demonstrate superior performance to frequentist analysis techniques (Chapter 4).

CHAPTER 2

OVERCOME THE LIMITATION OF PHENOME-WIDE ASSOCIATION STUDIES (PHEWAS): EXTENSION OF PHEWAS TO EFFICIENT AND ROBUST LARGE-SCALE ICD CODES ANALYSES

2.1 Introduction

The first successful Genome-Wide Association Study (GWAS) was published by Ozaki et al. (2002). Since then, GWAS has been used as a main tool to identify new genetic associations in many diseases. With the accumulation of DNA biobank linked with electronic medical records (EMR), large number of genetic-disease associations can be conducted in one study. One alternative approach to assess the genetic-disease associations is the Phenome-Wide Association Study (PheWAS), proposed by Denny et al. (2013) and has demonstrated reproducibility from known associations on GWAS catalog (Buniello et al. (2019)). Following the concept of "reverse-GWAS", PheWAS analyses scan through large number phenotypes with a given genetic variant. It is an extremely helpful tool when trying to build disease comorbidity networks.

Current practice for PheWAS analyses models the genetic-disease associations with logistic regression consisting binary disease outcomes and a genetic variant exposure of patients. In the EMR system, ICD billing codes remain the most commonly used phenotype outcomes to assess patient's disease status. According to World Health Organization (WHO), ICD 9 codes have been used from year 1900 till now. Currently there are around 13000 codes available. Starting October, 2015, ICD 10 codes have officially entered the health system. The new ICD 10 system carries around 68000 codes with more classification options to categorize diseases. However, the billing codes data are known to be noisy. One statistical assumption for logistic regression is clear classification of cases and controls population. Since 2010, several studies have been done with raw ICD 9 codes or combined ICD 9 codes (Neuraz et al. (2013); Hebbring et al. (2015)) without accurate definition of the controls population. Although showing some degree of reproducibility, violating the assumption of logistic regression can lead to biased results. In 2016, Denny et al. (2016) et al proposed to aggregate ICD 9 codes into "Phecodes" based on the similarities of the different ICD 9 codes. Phecodes have since become standard choice for phenotypes in PheWAS. To be classified as a "case" for a Phecode, the subject needs to have at least 2 or more ICD 9 codes on different days mapping to that specific Phecode. Since most of the efforts in clinical practice have been spent on trying to define true cases, the remaining population might not be true controls. Therefore, one exclusion criteria list was manually composed for each Phecode, trying to filter out the potential misclassification of the remaining non-case population. To be classified as a true control for a Phecode, the subject can not have any ICD 9 code mapping to the Phecode and any ICD 9 code in the exclusion criteria list for the Phecode.

From the clinical standpoint, manually-compiled exclusion criteria lists, considering disease comorbidities, might help improve accuracies of the regression estimates by filtering out misclassification in the non-case population. This procedure acts as Gold-standard. However, there are few drawbacks with this approach. First, the disease networks are large and complex. There are still plenty of unknown relationships between the diseases. It is highly possible that the complete disease networks weren't fully considered when compiling the exclusion criteria lists. Next, the accuracies of the lists cannot be guaranteed without extensive data curation process. The PheWAS analyses with Phecodes could give a better understanding of the general view of the disease structures but result in loss of information. The ICD codes can potentially provide more details on the disease status, treatment information and so on. The costly curation process in current PheWAS analyses limits the efficiency of large-scale analyses that take full advantage of all structured phenotypic information available in EMR.

To overcome this limitation in PheWAS analyses, we turned to the original study design. In case-control studies, the data are collected based on the disease status. In contrast, retrospective data are accessed after some patients have already developed the outcomes. The investigators then jump back in time to identify the exposure status at a point of time before any development of the disease outcome. Lastly, one can determine whether the subject subsequently develops the outcome. In PheWAS, we have one exposure, usually a SNP, with multiple disease outcomes. The proportions of subjects that are in exposed and unexposed groups are the same for each SNP-outcome analysis. Further, SNPs exist before the development of the general diseases in the system. With these two conditions, we can think of the population being divided by the exposure status and followed throughout time in the EMR system. This kind of study can be viewed as a retrospective cohort study. Once we established the study design, we further explored other options to assess the genetic-disease associations. In case-control studies, since the investigators already set up the disease prevalence, odds ratio remains the only measurement. In contrast, in retrospective cohort studies, relative risk is also a valid measurement. In our study, we evaluated the usage of relative risk as a measurement in PheWAS analyses.We demonstrated that relative risk, although not completely address the misclassification issues in EMR data, overcomes the need for clear classification of the true control population. Relative risk is a robust and efficient estimator that enables analyses on larger-scale ICD codes.

In Theoretical Background section, we demonstrate via theoretical formula on how relative risk can be free of biases due to misclassification. In Simulation section, we show the performance of relative risk comparing to Gold-standard odds ratio with several combinations of outcome prevalence and degrees of misclassification. In Real Data Analysis section, we illustrate with real data that relative risk model behaved similarly to Gold-standard model. Further, we generalize PheWAS analyses to ICD 9 codes. We were able to obtain additional relevant useful disease information that is missed in PheWAS analyses. Lastly, in Discussion section, we explore other available methods and discuss the implication of our work to the scientific field.

2.2 Theoretical Background

To begin with, we would like to explore different scenarios with theoretical formula derivation. Table 2.1 illustrates the distribution of hypothetical data with binary exposure (E) and binary outcome (Y). Letter a, b, c and d denote the number of observations in each E, Y combination. The sum of a, b, c, d equals the number of total observations (N).

Table 2.1: 2 × 2 table with Outcome and Exposure in hypothetical data

	Y = 1	Y = 0
E = 1	a	b
E = 0	c	d

With Table 2.1, the formula for odds ratio is $\frac{ad}{bc}$ and $\frac{\frac{a}{(a+b)}}{\frac{c}{c+d}}$ for relative risk. The odds ratio can be converted to relative risk with formula (Grant (2014)), $\frac{\text{OR}}{1-\text{p\_risk}+(\text{p\_risk}\times\text{OR})}$ where p_risk denotes the risk in the control groups (P(Y=1|E=0)), in this case, $\frac{c}{c+d}$. There are two scenarios that could happen when defining cases and controls: cases misclassified into controls and controls misclassified into cases. For the purpose of PheWAS analyses, the scenario where cases are misclassified into controls is the primary focus and shown in the main text. The method and simulation result for the scenario where controls are misclassified into cases is shown in Appendix. We can derive the formula when cases were misclassified into controls with Table 2.2. We denote Z as the proportion of cases being misclassified as controls and T as the proportion of the misclassification **not** being removed.

Table 2.2: 2 × 2 table with Outcome and Exposure; cases misclassified as controls

	Y = 1	Y = 0
E = 1	a - aTZ	b + aTZ
E = 0	c - cTZ	d + cTZ

The formula for odds ratio becomes $\frac{ad+acTZ}{bc+acTZ}$. When misclassification is completely removed (T=0), odds ratio is $\frac{ad}{bc}$, unbiased and free of Z. When the estimated unbiased odds ratio is converted to relative risk, the estimated relative risk is $\frac{ac(1-Z)+ad}{bc+ac(1-Z)}$, a function of Z. When misclassification is completely ignored (T=1), odds ratio is $\frac{ac(1-Z)+ad}{bc+ac(1-Z)}$, biased and a function of Z. The bias increases as Z increases. It's worth noting that when Z approaches 1, the biased odds ratio approaches $\frac{\frac{a}{(a+b)}}{\frac{c}{c+d}}$, the relative risk. When the estimated biased odds ratio is converted to relative risk, the estimated relative risk is $\frac{\frac{a}{(a+b)}}{\frac{c}{c+d}}$, unbiased and free of Z. For the direct estimation of relative risk, with the goal of our paper, we do not remove any observation (T=1). The estimated relative risk is $\frac{\frac{a}{(a+b)}}{\frac{c}{c+d}}$, unbiased and free of Z.

With the theoretical formula, the estimated odds ratios are unbiased only when all misclassification is removed (T=0). The direct estimation of relative risk or conversion from biased odds ratios where no observation was removed in the original dataset (T=1) produce unbiased relative risk.

2.3 Methods

2.3.1 Data

2.3.1.1 Univariate simulation

To validate the theoretical formula, we first conducted univariate simulation. In univariate simulation, we included only a binary exposure (E) as a predictor and an binary outcome (Y). We denote proportion of cases being misclassified as controls as Z. The probability of exposure (Pe) was set to 0.1. The prevalence of Y (P1) includes 0.02 and 0.3 to explore the influence of the prevalence on the estimates. The true relative risks (RR) were set to 0.5 and 3. Four P1 and RR combinations are listed as below:

- P1 = 0.02 and RR = 0.5
- P1 = 0.3 and RR = 0.5
- P1 = 0.02 and RR = 3

- P1 = 0.3 and RR = 3

To explore the influence of varying levels of misclassification (Z), in each P1 and RR combination, 5 conditions of Z were implemented, including 0 (No misclassification), 0.1, 0.25, 0.5 and 0.75 (75% of the true cases misclassified into controls). 500 simulations and 7000 observations (N) were conducted for each P1 and RR combination.

The distribution for binary E follows Binomial(N, 0.1). The conditional probabilities, P(Y=1|E=1) and P(Y=1|E=0) were calculated with pre-specified Pe, P1 and RR according to the above list. With Bayes rules, the conditional probabilities can be computed with known joint probabilities with the relative risk formula where P(Y=1, E=1) equals $\frac{\text{RR}\times\text{P1}\times\text{Pe}}{(\text{RR}\times\text{Pe})-\text{Pe}+1}$ and P(Y=1, E=0) equals P1 - P(Y=1, E=1). Lastly, P(Y=1|E=1) equals P(Y=1, E=1)/Pe and P(Y=1|E=0) equals P(Y=1, E=0)/(1-Pe). Y was simulated according to the conditional probabilities, P(Y=1|E=1) and P(Y=1|E=0). By directly calculating the conditional probabilities, we can easily make sure all the 4 conditional probabilities, P(Y=1|E=1), P(Y=1|E=0), P(Y=0|E=1) and P(Y=0|E=0), are positive. After E and Y were simulated, Z proportion of cases (Y=1) was randomly converted to controls (Y=0). For the additional simulation where controls are misclassified as cases in Appendix, the simulation procedures follow the above univariate simulation with P1 = 0.3 and RR = 0.5. After E and Y were simulated, Z proportion of controls (Y=0) was randomly converted to cases (Y=1).

2.3.1.2 Multivariable simulation

We included a multivariable simulation mimicking the commonly-used additive modeling of SNPs for the exposure variable and the presence of a continuous covariate (C). The data were simulated with a log-binomial model: log(P(Y=1|E, C)) = log(0.15) + log(2)E + log(1.3)C. The exposure variable E was simulated with categories 0, 1, 2 indicating homozygous dominance, heterozygotes and homozygous recessive genotypes. The corresponding probabilities were 0.6, 0.3 and 0.1. The covariate C was simulated with Normal distribution, mean = 0 and variance = 0.3. The prevalence of Y was approximately 0.456, empirically. Similar to univariate simulation, Z proportion of cases (Y=1) was randomly converted to controls (Y=0).

2.3.1.3 Clinical data

The data were obtained from Vanderbilt BioVU database. The SNPs were selected from the previous published paper (Denny et al. (2013)) that passed the quality control after imputation. The four SNPs are s660895, rs1847134, rs258322 and rs4977574.

For PheWAS analyses, there are 44764 subjects available from the BioVU cohort. For Phecodes, following the procedure in Denny et al. (2013), we excluded Phecodes with case number < 25. 1446 Phecodes are available for analyses. The exclusion criteria will be implemented for Gold-standard logistic regression following the the published criteria in Denny et al. (2013). For ICD 9 analysis, there are 44846 subjects available from the BioVU cohort. For the ICD 9 codes, we excluded ICD 9 codes with case number < 25. 6329 ICD 9 codes are available for analyses.

2.3.2 Statistical analysis

2.3.2.1 Simulation study

For each simulation, we included three models for relative risk comparison, bias-corrected (Firth (1993)) Poisson, Logistic regression with exclusion criteria and Logistic regression without exclusion criteria. For the logistic regression models, the odds ratios were converted to relative risk with the conversion formula presented in Theoretical Background section. We also reported comparison of Logistic regression with exclusion criteria and Logistic regression without exclusion criteria for odds ratios. Logistic regression with exclusion criteria is the current practice of PheWAS, also referred to as the Gold-standard model. Logistic regression was conducted on dataset that the misclassification is manually removed. In our simulation, we assumed that the Gold-standard model can remove all misclassification (T=0). Logistic regression without exclusion criteria refers to conducting logistic regression on dataset without removing any observation (T=1). For Poisson model, the regression was also conducted on the dataset without removing any observation.

2.3.2.2 Clinical application

To illustrate the potential biases in BioVU data, Logistic regression with and without exclusion criteria were performed with SNP rs14483486 in Figure 2.4 following a table demonstrating when the biases are the most obvious. The percent biases were calculated with the odds ratios estimated from the Logistic regression model with and without exclusion criteria (100×(OR_without - OR_with)/OR_with). The prevalence and misclassification rate were calculated assuming the observations removed after exclusion criteria were true cases. An example calculation of the prevalence and misclassification rate is given in Clinical Application Results section.

Further, to compare the performance of Logistic regression with exclusion criteria (Gold-standard model) and the proposed Poisson model, we conducted analyses with the four SNPs previously published in the PheWAS papers (Denny et al. (2013)), rs660895 (Figure 2.5), rs1847134 (Appendix, Figure 2.8), rs258322 (Appendix, Figure 2.9) and rs4977574 (Appendix, Figure 2.10). In each figure, we illustrate the results from Gold-standard logistic regression, bias-corrected Poisson model and Poisson model with robust sandwich standard errors (Stock and Watson (2008)). Note that currently, the robust standard errors option is not always compatible with the bias-correction option (especially when the correction is most needed, outcomes with very low prevalence). In our previous simulation studies (not shown), the bias-correction procedure only affects codes with infinite estimates, which constitutes less than 0.5 percent of the codes in real data. We controlled the statistical significance at p-values $\leq 1.20 \text{ x } 10^{-5}$ which equals to controlling false discovery rate (FDR) < 0.01 level for all 4 PheWAS analyses combined (5784 comparisons) or Bonferroni correction at α around 0.02 for 1 single PheWAS study (1446 comparisons). Lastly, we implemented the proposed Poisson model with robust sandwich standard errors on ICD 9 codes with the same SNPs used in PheWAS analyses. We controlled the statistical significance at p-values $\leq 4.16 \text{ x } 10^{-6}$ which equals to controlling FDR < 0.01 level for all 4 ICD 9 analyses combined (25316 comparisons) or Bonferroni correction at α around 0.03 for 1 single ICD 9 study (6329 comparisons). Results for ICD 9 analyses are shown in Figure 2.6, Figure 2.11 (Appendix), Figure 2.12 (Appendix) and Figure 2.13 (Appendix). The significance criteria chosen here might be slightly conservative for 4 studies. However, our criteria are comparable to criteria used in Denny et al. (2013). The choice of the criteria should not influence the results for methods comparison purposes. Age, gender and the first three genetic principle components were included as covariates for PheWAS and ICD 9 analyses.

2.4 Results

2.4.1 Simulation study

Following the formula in Theoretical Background section, when cases are misclassified into controls, odds ratio from Logistic regression with exclusion criteria stays unbiased with varying degree of misclassification. When misclassification is not completely removed, odds ratio is biased and a function of Z. As misclassification rate increases, biased odds ratio approaches the true relative risk. Therefore, if the true odds ratio and true relative risk differ more, the biases should become more profound.

The univariate simulation results are shown in Figure 2.1 and Figure 2.2 with boxplots comparing the proportion of cases misclassified as controls (X axis) and the

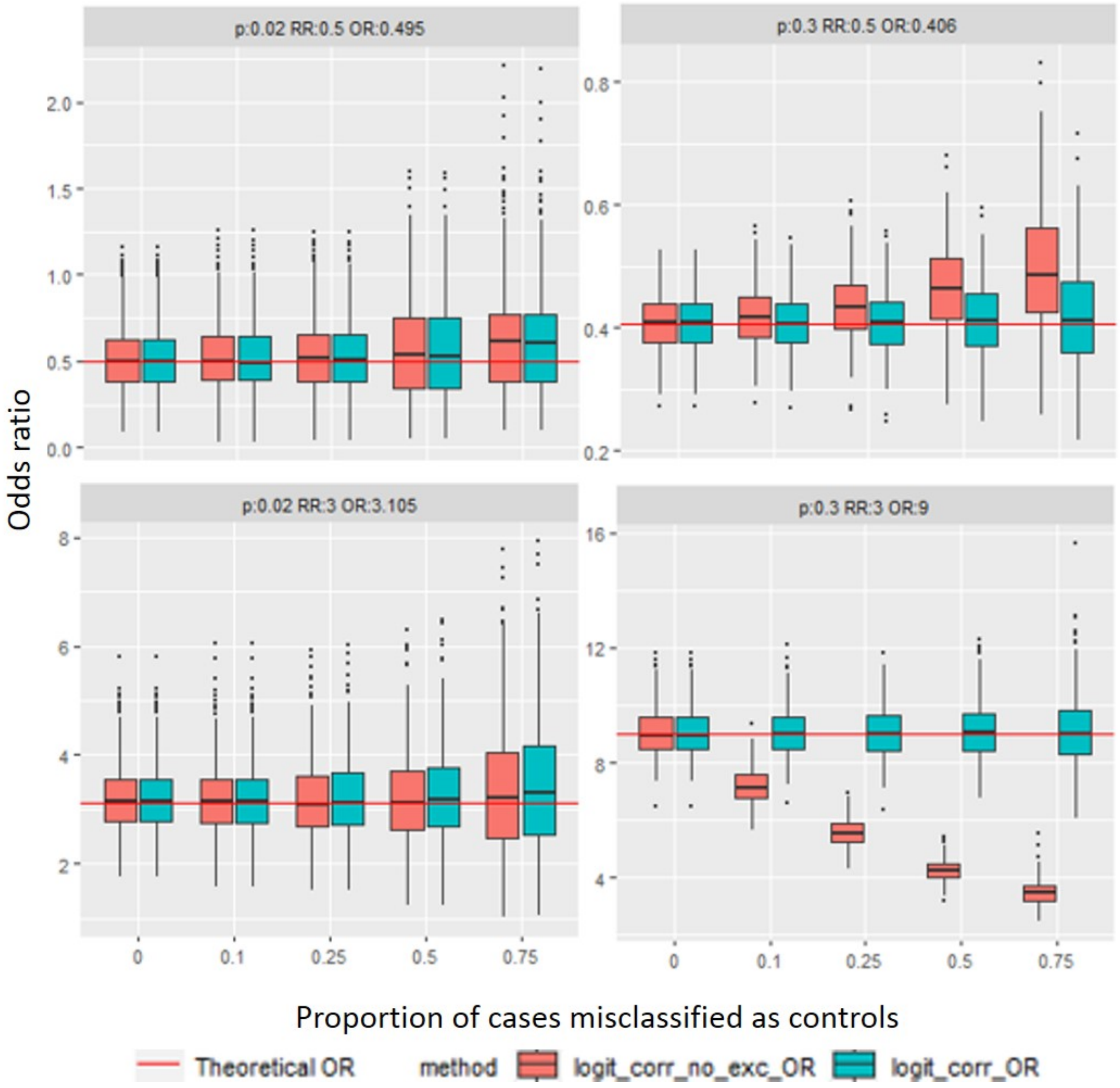


Figure 2.1: Univariate simulation result: Odds ratio boxplots for prevalence = 0.02, 0.3 and exposure probability = 0.1 with varying true odds ratios. The true odds ratios are 0.495 (top left), 0.406 (top right), 3.105 (bottom left) and 9 (bottom right). The x-axis denotes the proportion of cases being misclassified as controls. The y-axis denotes the odds ratio estimates from the logistic regression with (teal boxes) and without exclusion criteria (red boxes).

estimates for odds ratios/relative risk (Y axis). As shown in Figure 2.1, when the prevalence is 0.02, the difference between the true relative risk and the true odds ratio is small (0.5 versus 0.495 and 3 versus 3.105). As Z increases, biases in odds ratios are almost negligible when ignoring the misclassification. However, when the prevalence becomes 0.3, the true relative risk and odds ratio differ more (0.5 versus 0.406 and 3 versus 9), the biases becomes noticeable as Z increases. The biased odds ratios bias toward the true relative risk when ignoring the misclassification. From

Figure 2.1, unbiased odds ratios can only be obtained with logistic regression when exclusion criteria were applied (all misclassification is removed).

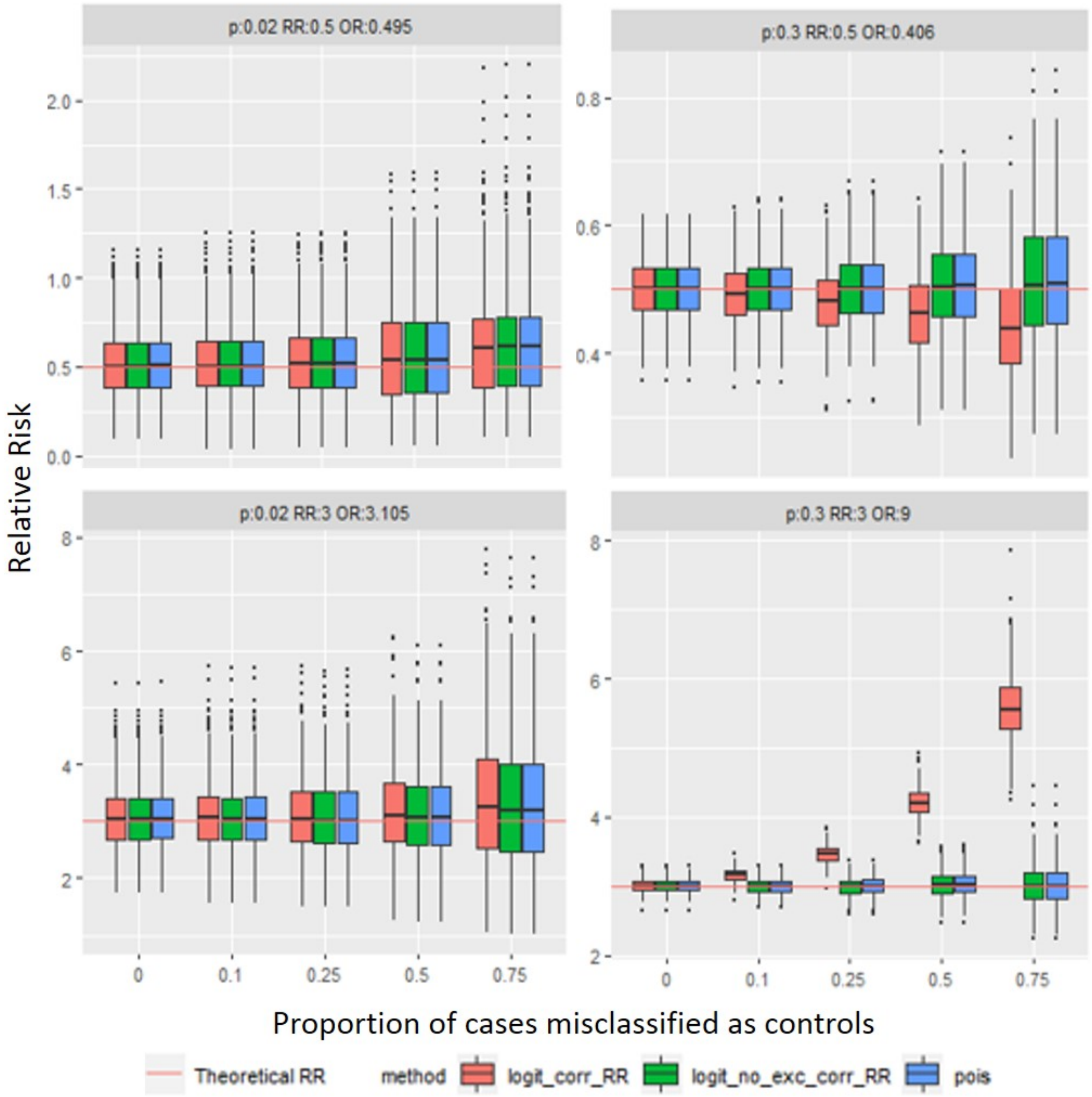


Figure 2.2: Relative risk boxplots for prevalence = 0.02, 0.3 and exposure probability = 0.1 with varying true relative risk ratios. The relative risk ratios are 0.5 (top row) and 3 (bottom row). The x-axis denotes the proportion of cases being misclassified as controls. The y-axis denotes the relative risk estimates from Poisson model (blue boxes), converted relative risk estimates from logistic regression with (red boxes) and without exclusion criteria (green boxes).

Counter-intuitively, when the estimated unbiased odds ratios were converted to relative risk, the estimates were biased. The magnitudes of biases was shown as a function of Z in Theoretical Background section. The corresponding simulation result is shown in Figure 2.2. Relative risk converted from unbiased odds ratio from Logistic regression with exclusion criteria is biased. Similar to Figure 2.1, the biases are not noticeable when the prevalence of Y is smaller. As the prevalence increases to 0.3, the bias becomes profound as Z increases. The relative risk estimated from Poisson model and converted from biased odds ratio stay unbiased in all scenarios. The univariate simulation result confirms the conclusion obtained in Theoretical Background section.

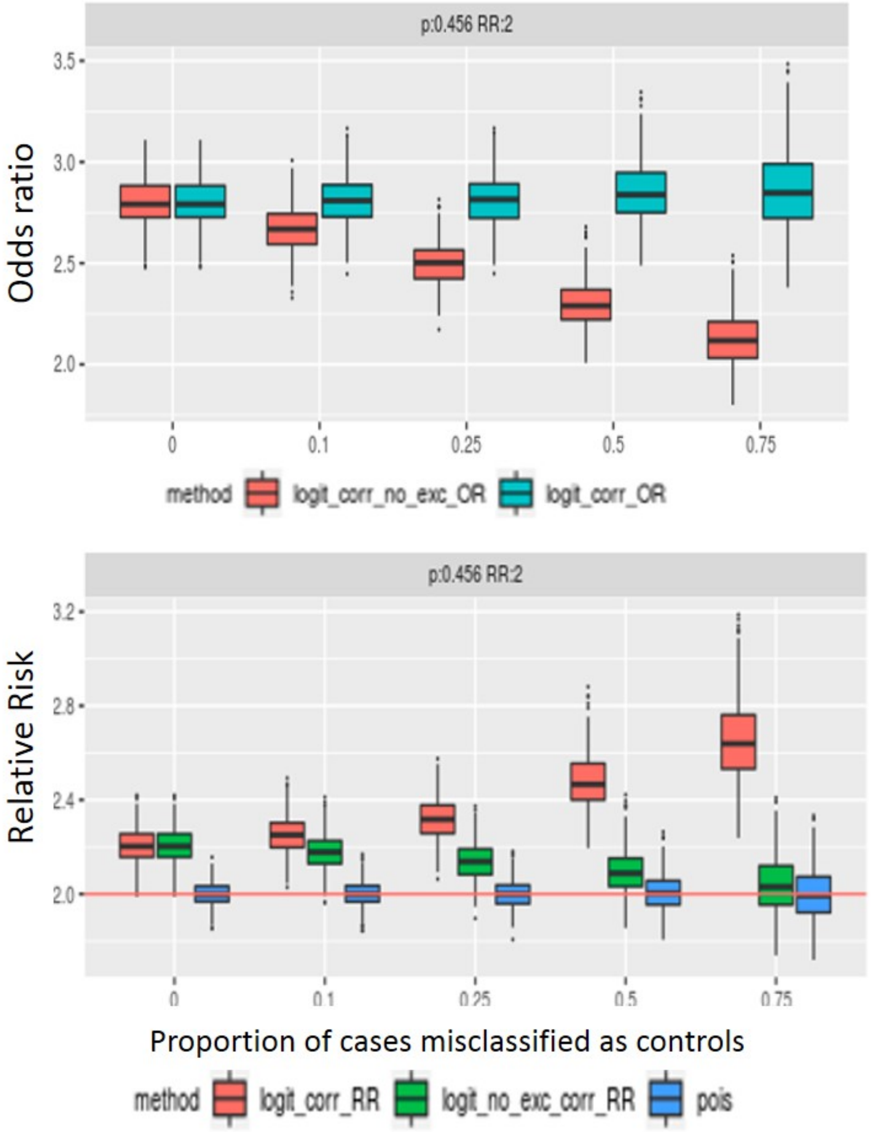


Figure 2.3: Odds ratios boxplot (top) and relative risk boxplot (bottom) for exposure probability = 0.1, prevalence = 0.456 and true relative risk = 2. The x-axis denotes the proportion of cases being misclassified as controls. The y-axis for the top figure denotes the odds ratio estimates from the logistic regression with (teal boxes) and without exclusion criteria (red boxes). The y-axis for the bottom plot denotes the relative risk estimates from Poisson model (blue boxes), converted relative risk estimates from logistic regression with (red boxes) and without exclusion criteria (green boxes).

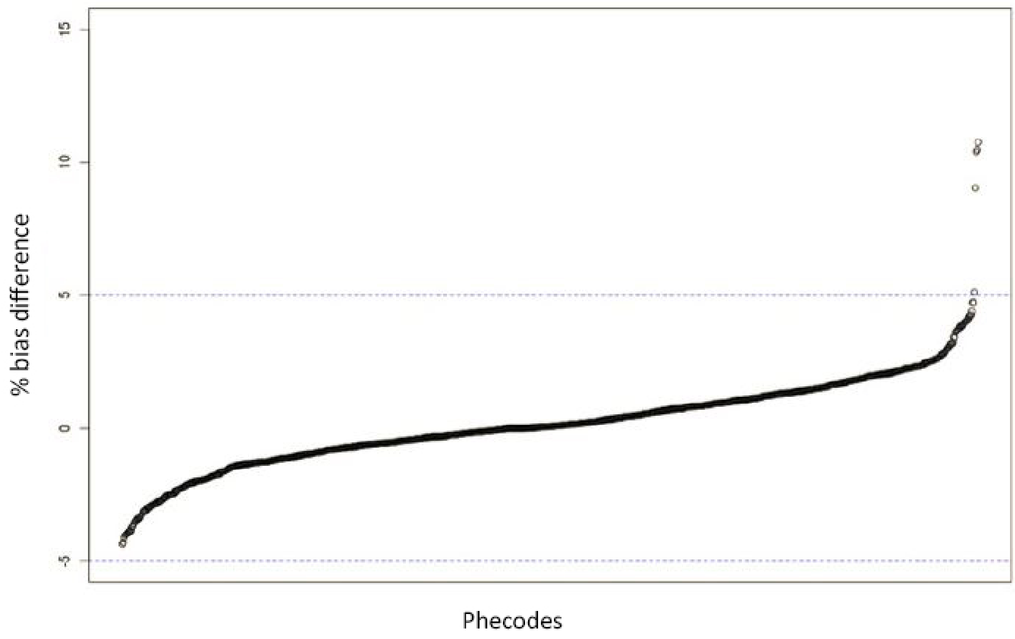
The multivariable simulation result is shown in Figure 2.3. With the additive modeling of exposure and the presence of covariate C, the odds ratios (Figure 2.3, top) are unbiased only when all misclassification is removed (T=0), same as what we observe in univariate simulation. In the bottom plot, the estimated relative risk converted from unbiased odds ratio is biased and increases as Z increases. Unbiased relative risk is only obtained in the Poisson model. Interestingly, the relative risk converted from biased odds ratio is biased, unlike what we observe in Figure 2.2. We hypothesize that the reason for the biases might be due to the added covariate C . The conversion formula presented in Theoretical Background section does not consider the presence of any additional covariate.

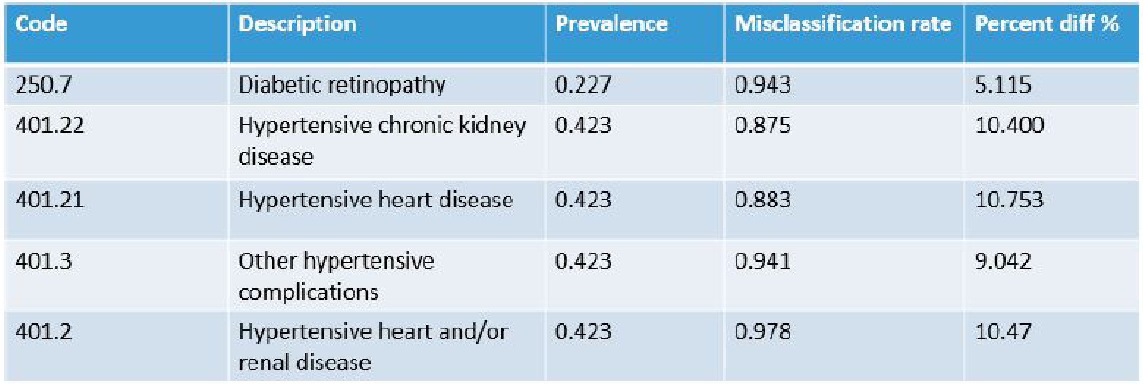
Although not directly pertaining to PheWAS analyses, the scenario where controls were misclassified as cases is shown in Appendix. In Appendix, the formula for both relative risk and odds ratio obtained ignoring misclassification are biased and functions of Z. In odds ratio estimation (Appendix, Figure 2.7, left), exclusion criteria are needed to obtain unbiased odds ratios. In relative risk estimation (Appendix, Figure 2.7, right), all methods failed to obtain unbiased relative risk.

2.4.2 Clinical application

In the Simulation Results section, biases mainly occurred when the prevalence of the outcome was high. The biases increased as misclassification rates increased. In Figure 2.4, with SNP rs14483486 as an example, Phecodes with more than 5% biases include common diseases like diabetic and hypertension related diseases with prevalence > 0.2. In addition, The misclassification rates for these Phecodes are high, more than 0.85. The observations match the trend in the simulation results. Since we do not know the true classification of the control population, the prevalence and misclassification rates were calculated assuming the observations removed after exclusion criteria were true cases, misclassified as controls in the original data. Take Phecode 250.7 for example, there were 34592 subjects classified as true controls, 574 as true cases and 9598 observations removed. We assumed that subjects removed were true cases. The estimated prevalence was $(574 + 9598)/44764 = 0.227$. The estimated misclassification rate was $9598/(9598+574) = 0.943$.

To compare the performance of Poisson models and Gold-standard model, we conducted PheWAS analyses on the BioVU data with 4 SNPs. In general, Poisson models obtain comparable inference to Gold-standard model. Gold-standard model captures a large number of disease-genetic associations including Type I diabetes and Rheumatoid arthritis related Phecodes with SNP rs660895 (Figure 2.5, top). Both bias-uncorrected Poisson model with robust standard errors (Figure 2.5, bottom) and





Code	Description	Prevalence	Misclassification rate	Percent diff %
250.7	Diabetic retinopathy	0.227	0.943	5.115
401.22	Hypertensive chronic kidney disease	0.423	0.875	10.400
401.21	Hypertensive heart disease	0.423	0.883	10.753
401.3	Other hypertensive complications	0.423	0.941	9.042
401.2	Hypertensive heart and/or renal disease	0.423	0.978	10.47

Figure 2.4: Bias illustration with SNP rs14483486. In the top figure, the y-axis denotes the percent difference of the odds ratios estimated from univariate logistic regression with and without exclusion criteria. We denote this as percent bias as the odds ratios from the Gold-standard model acts as the truth. The percent bias was calculated with formula: (100×(OR_without - OR_with)/OR_with). Each dot represents a Phecode. The bottom table summarizes information, including the Phecode ID, description of the Phecode, prevalence of the Phecode in the study population and misclassification rates (estimated from the manually compiled exclusion criteria list) of Phecodes with greater than 5% absolute bias.

the bias-corrected Poisson model (Figure 2.5, middle) are able to obtain all the significant Phecodes captured in Gold-standard method. In Figure 2.8 (Appendix), Gold-standard model only captures one Phecode, "Other non-epithelial cancer of skin" with SNP rs1847134 (Appendix, Figure 2.8, top). Both bias-uncorrected Poisson model with robust standard errors (Appendix, Figure 2.8, bottom) and bias-corrected Poisson model (Appendix, Figure 2.8, middle) are able to catch the Phecode. In Figure 2.9

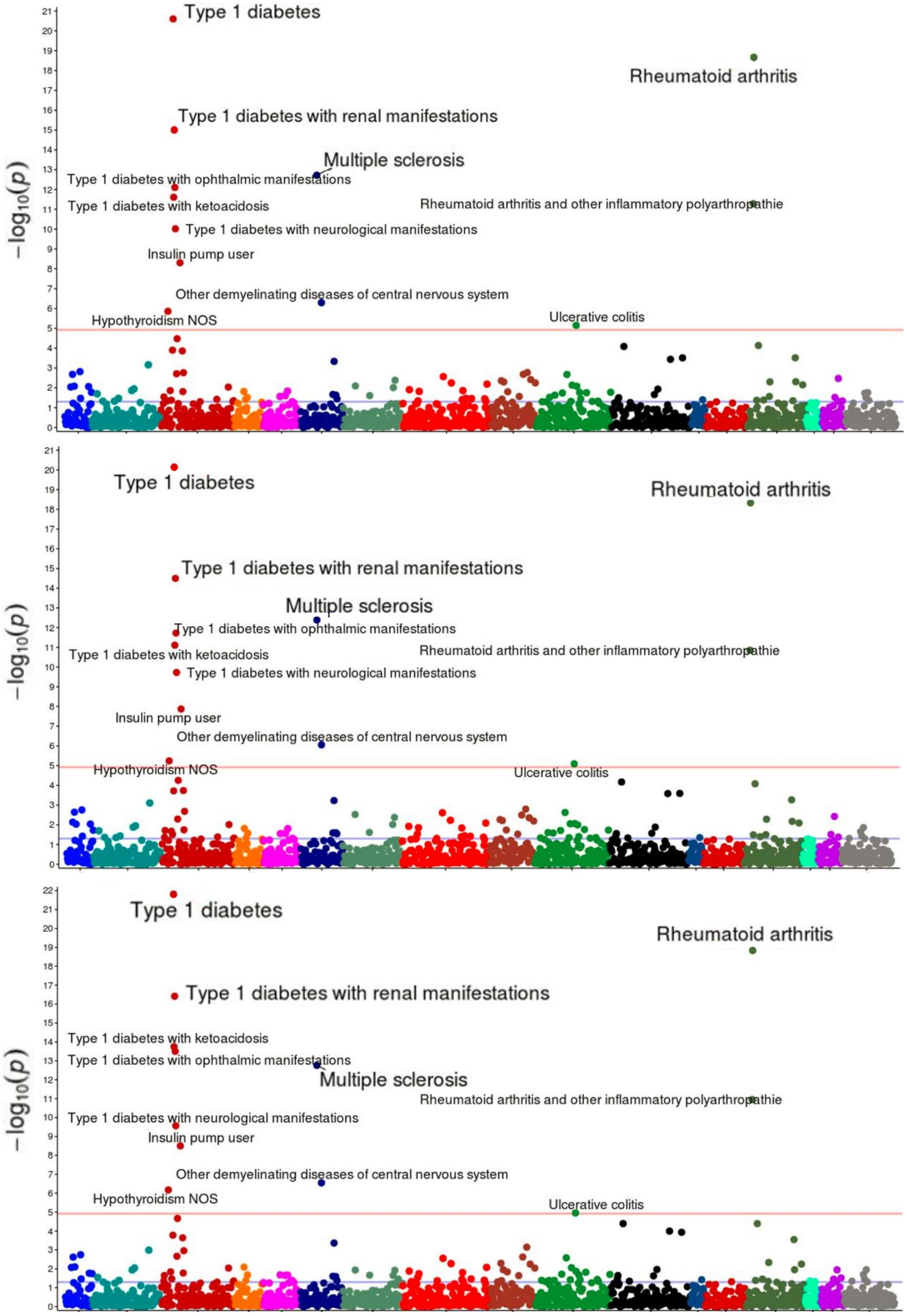


Figure 2.5: PheWAS analysis result for SNP rs660895 with Gold-standard Logistic regression model (top), bias-corrected Poisson (middle) and robust Poisson model (bottom). Each dot represents one Phecode. The red line indicates the significance line. X axis denotes the negative log 10 p-values. Top selected codes are Type I diabetes, Rheumatoid arthritis, Type I diabetes with renal manifestations and Multiple sclerosis. The maximum value of X axis is 21 for the top two plots and 22 for the bottom plot.")

(Appendix), although both Poisson models fail to get the Phecode, "Neoplasm of uncertain behavior of skin", captured by the Gold-standard model with SNP rs258322

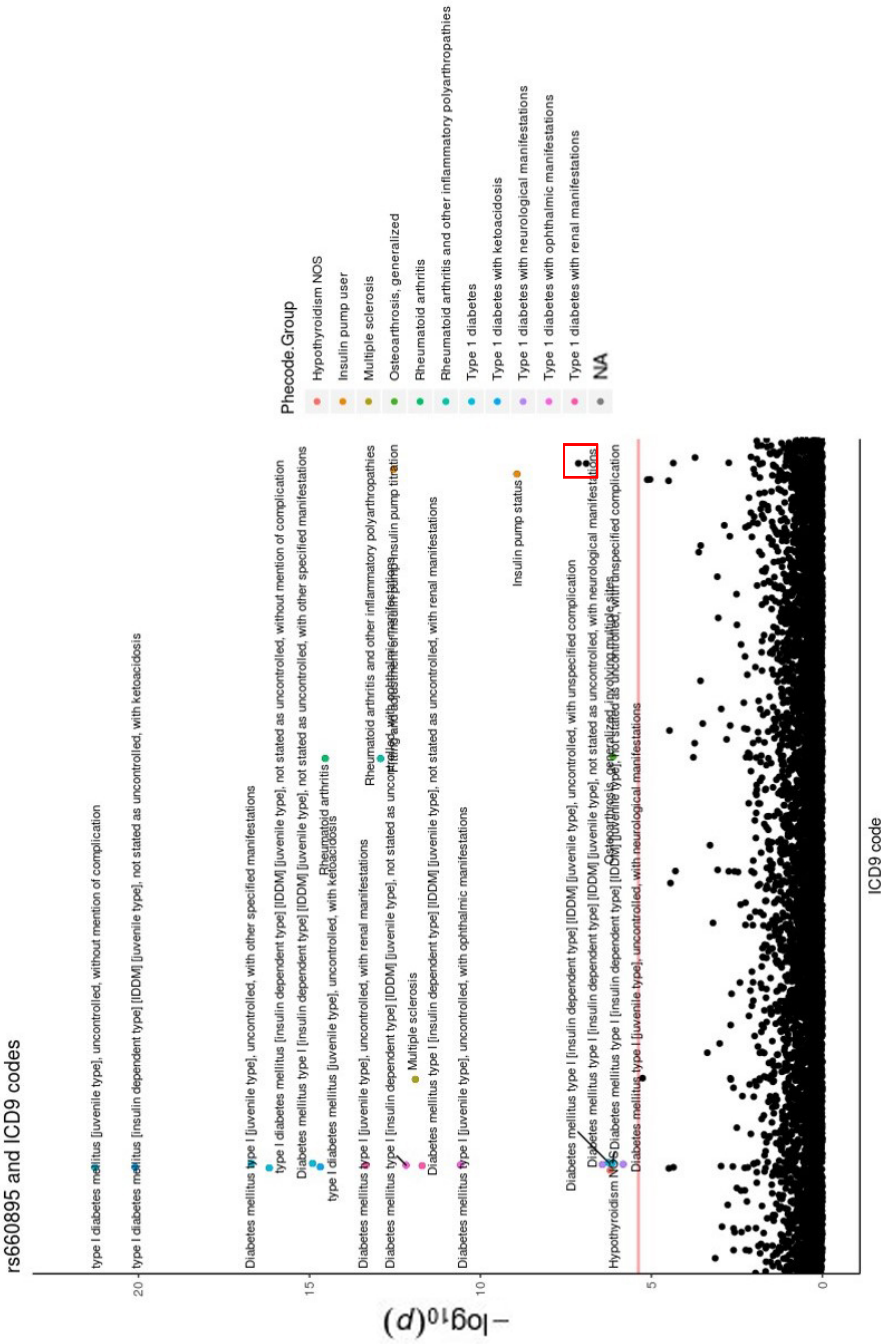


Figure 2.6: Result for ICD 9 analysis with SNP rs660895. Each dot represents an ICD 9 code. Y-axis denotes the negative log 10 P-values. The colors of the dot corresponds to Phecode groups indicating in the legend. The red line indicates the significance line. The group NA means that the codes do not correspond to any Phecode group. The significant ICD 9 codes in the red boxes correspond to V42.83, Pancreas replaced by transplant and V58.64, Long-term usage of non-steroidal anti-inflammatory. These two codes do not belong to any Phecode group.")

(Appendix, Figure 2.9, top), they are able to obtain all other skin cancer related Phecodes (Appendix, Figure 2.9, middle and bottom). In Figure 2.10 (Appendix), bias-uncorrected Poisson model with robust standard errors (Appendix, Figure 2.10, bottom) is able to catch all important Phecodes captured in Gold-standard model about Coronary artery disease (Appendix, Figure 2.10, top). the Phecode, "Unstable angina", close to the significance borderline, is missed by bias-corrected Poisson model (Appendix, Figure 2.10, middle).

Since the estimation of relative risk can bypass the need for exclusion criteria lists, we further extend Poisson model to large-scale phenomes. We conducted analyses on ICD 9 codes via bias-uncorrected Poisson model with robust standard errors. The corresponding ICD 9 codes analysis result for Figure 2.8 (Appendix) is shown in Figure 2.11 (Appendix). The ICD 9 codes obtained belong to non-epithelial skin cancer related Phecode groups, "Other non-epithelial cancer of skin" and "Neoplasm of uncertain behavior of skin", similar to the result obtained in PheWAS analysis (Appendix, Figure 2.8). The corresponding ICD 9 codes analysis result for Figure 2.9 (Appendix) is shown in Figure 2.12 (Appendix). The ICD 9 codes obtained belong to epithelial skin cancer and skin cancer related to sun exposure Phecode groups, similar to the result obtained in PheWAS analysis (Appendix, Figure 2.9). The corresponding ICD 9 codes analysis result for Figure 2.10 (Appendix) is shown in Figure 2.13 (Appendix). The ICD 9 codes obtained belong to Coronary artery diseases Phecode groups, including "Coronary atherosclerosis","Angina pectoris" and "Unstable angina", similar to the result obtained in PheWAS analysis (Appendix, Figure 2.10). These results show that with ICD 9 codes analyses, we are able to discern the major related diseases as in PheWAS analyses while obtaining more detailed description of the diseases. In addition, with ICD 9 codes analyses, we are able to capture information that might not be available in Phecode groupings. Take SNP rs660895 for example, the corresponding ICD 9 codes analysis result is shown in Figure 2.6. All but two ICD 9 codes obtained belong to the Phecode groups captured in PheWAS analysis (Figure 2.5). ICD 9 analysis captures 2 additional ICD 9 codes that were not previously defined in any Phecode group, denoting as "NA" (Figure 2.6). These two ICD 9 codes are V42.83 (Pancreas replaced by transplant) and V58.64 (Long-term (current) use of non-steroidal anti-inflammatories). Based on the descriptions, these two ICD 9 codes relate to the surgical and drug-usage aspects of diabetes which are highly correlated with the Type I diabetes related Phecodes captured in Figure 2.5.

2.5 Discussion

PheWAS analyses were designed to provide quick scans to assess disease-genetic associations in the EMR data. Logistic regression is commonly used to model binary disease outcomes. In our setting, the disease outcomes are Phecodes, the aggregation of ICD 9 billing codes. To correctly implement logistic regression, strict assumption on the accuracies of the defined true "case" and "control" labels is required. It is our belief that more cares have been given to defining a true case and misclassification happens mainly in the remaining non-case population. Efforts have been made in Denny et al. (2013) to manually compile exclusion criteria lists to classify true controls for unbiased odds ratio estimates. Denny et al. (2013) was able to replicate several SNP-disease relationships listed in GWAS catalog. However, the manual compilation of the exclusion criteria lists has several drawbacks, including long data curation time and subjectivity of the lists. Treating Phecodes as disease outcomes, we also lose precision during the code aggregation. ICD 9 codes contain more detailed information about a specific disease. Nevertheless, the limitation of manual compilation procedure hinders the extension of PheWAS to larger-scale ICD codes analyses.

From the statistical standpoint, this issue of classifying non-case population can be viewed as outcome misclassification in logistic regression. Different methods have been proposed to address this issue including validation data-based methods and model-based methods. In the validation data-based methods, the investigator could have part of the data validated externally and obtained better model sensitivity and specificity with these validated data (Edwards et al. (2013); Lyles et al. (2011)). In PheWAS analyses, it is not easy to have validated data in the EMR system due to the large number of outcomes being compared. Further, if it takes similar time to obtain the validation dataset as compiling exclusion criteria, there's no advantage of the validation data-based methods over the exclusion criteria list procedure. The model-based methods incorporate the misclassification rate as a parameter in the models. By adjusting for the misclassification rate, the coefficient estimate for the exposure variable should be unbiased. Liu and Zhang (2017) proposed to incorporate misclassification rate parameter in the model and estimate the parameter with Fisher scoring algorithm. The simulation demonstrated that with the correct model specification, the odds ratio estimates are within 5% biases range. In addition, the method does not require any additional validation data. However, the algorithm doesn't work well in higher misclassification rates scenarios with potentially outputting invalid standard deviation estimates. In their simulation studies, the misclassification rates were only tested to 0.2. In PheWAS studies, the misclassification rates could range from 0.01

to 0.97. Therefore, more work need to be done to accommodate scenarios with larger misclassification rates.

In this study, we evaluated the usage of relative risk as a valid estimator in the EMR setting for binary outcomes, accounting for potential misclassification in the non-case population. From Simulation section, we observed that biases occurred when misclassification in the non-case population was ignored when estimating odds ratios. The biases increased as outcome prevalence and misclassification rates increased. The relative risk estimates were unbiased in all scenarios without the need for clear definition of true controls population. Both univariate and multivariable simulation results confirmed the theoretical formula we derived. In reality, true effect size estimates and misclassification rates for disease codes are unknown and can vary widely. Without taking into account of the misclassification, the effect size estimates are inaccurate which can lead to misleading interpretations and inference of the genetic-disease relationships.

When Poisson models were implemented on real data, the performance was similar to Gold-standard model. Theoretically, when Poisson model is implemented on a binary outcome, it tends to be more conservative than logistic regression. One common alternative model to estimate relative risk is log-binomial model. Since log-binomial distribution is a log-transformation of the distribution in logistic regression, the variance should be correct. However, log-binomial model has been reported to behave poorly and constantly fail to converge under certain settings, especially when the prevalence of the outcomes is low (Williamson et al. (2013); Marschner and Gillett (2012)). Fortunately, the over-conservative issue in Poisson model can be overcomed by applying robust sandwich standard errors (Stock and Watson (2008)) to obtain the correct variance. This can be seen in Figure 2.10 (Appendix) where the Phecodes captured by Gold-standard model were all captured by Poisson model with robust sandwich standard errors and not entirely by the non-robust Poisson model. Another well-known issue of modeling binary outcomes is small cell count which can lead to infinite effect size estimates. The infinite effect size estimates usually arise when one of the cell in a 2 × 2 table is 0. This issue can be corrected by implementing Firth correction (Firth (1993)). However, the Firth correction option is not always compatible with the robust sandwich standard errors. Therefore, following the preprocessing step used by Denny et al. (2013), we have excluded codes with less than 25 cases to reduce the need for Firth correction. In general, we recommended to apply robust sandwich standard errors with Poisson model to obtain better inference if possible.

In ICD 9 analyses, ICD 9 codes captured by Poisson model with robust sandwich standard errors match with Phecodes captured in the corresponding PheWAS analyses. In reality, ICD 9 analyses are blinded with Phecode grouping information. This indicates that the model has enough power to obtain similar inference in ICD 9 analyses as in PheWAS analyses. In addition to capturing the important codes that match with their Phecode groups, more than 1 ICD 9 codes could be captured for 1 Phecode group. This indicates that we are able to learn what conditions of the disease contribute most in a certain disease group. Phecodes are aggregation of billing codes and not all ICD 9 codes were used to define Phecodes. Some information is lost in PheWAS analyses. In Figure 2.6, 2 codes are captured with no corresponding Phecode groups. These 2 codes contain information that are tightly related to other disease codes captured in the same analysis. Code V42.83 relates to Pancreas replaced by transplant. This is a surgery code that relates to diabetes, one of the main diseases captured in the SNP-disease analysis. Code V58.64 relates to the long-term usage of non-steroidal anti-inflammatory, a pain reliever that has been used mostly for patients with arthritis pain (Crofford (2013); Wongrakpanich et al. (2018)). This code is highly related to several arthritis codes captured. The additional information obtained from ICD 9 analysis is associated with the treatments and procedures the patients have undergone and could be very useful from clinical perspective.

In this study, we evaluated relative risk as an alternative estimator to assess SNP-disease associations that overcomes the misclassification issues in disease codes without clear definition of the controls population. Although several methods have been proposed to address the misclassification issue when estimating odds ratios, additional difficulties still exist in the context of EMR data. It is noteworthy to mention that using relative risk as an estimator does not completely solve the misclassification issue in the disease codes. Instead, it is a robust estimator that provides unbiased estimates without considering the misclassification in the control group. Statistical models that estimate relative risk are implemented in almost every existing software and ready to use. From our study, relative risk is a robust estimator that efficiently extend PheWAS analyses to larger-scale, phenome construction agnostic analyses of phenotypes (via ICD 9/10) and obtain additional clinical information that might not be captured with Phecodes.

2.6 Appendix

2.6.1 Theoretical Background: controls misclassified into cases

Table 2.3: 2 × 2 table with Outcome and Exposure; controls misclassified

	Y = 1	Y = 0
E = 1	a + bTZ	b - bZ
E = 0	c + dTZ	d - dZ

We can denote the odds ratio estimate as $\frac{ad+bdTZ}{bc+bdTZ}$ and the relative risk estimate as $\frac{\frac{a+bZ}{(a+b)}}{\frac{c+dz}{c+d}}$. Both the odds ratio and relative risk are functions of Z.

2.6.2 Figures

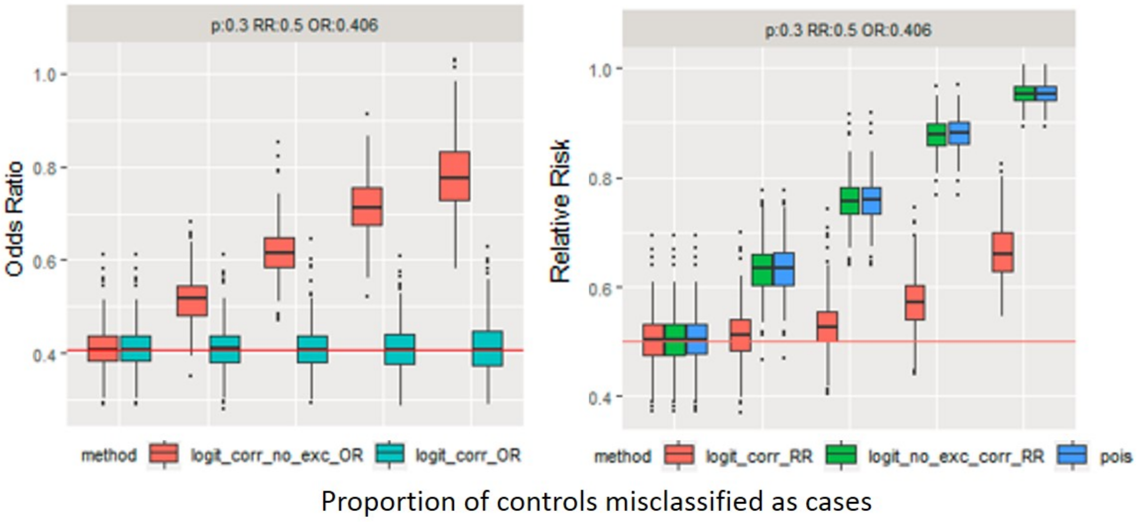


Figure 2.7: Univariate simulation result for controls misclassified as cases for odds ratios (left) and relative risk ratios (right). The prevalence = 0.3 and the exposure probability = 0.1. The true relative risk = 0.5 and the corresponding odds ratio = 0.406. The x-axis denotes the proportion of controls being misclassified as cases. the y-axis for the left plot denotes the odds ratios and relative risk estimates for the right plot. For the left plot, the red boxes correspond to odds ratios estimated from logistic regression without exclusion criteria. The teal boxes correspond to odds ratio estimated from logistic regression with exclusion criteria. For the right plot, the red boxes correspond to the converted relative risk estimates from logistic regression with exclusion criteria. The green boxes correspond to the converted odds ratio from the logistic regression without exclusion criteria. The blue boxes correspond to the relative risk estimated from Poisson model

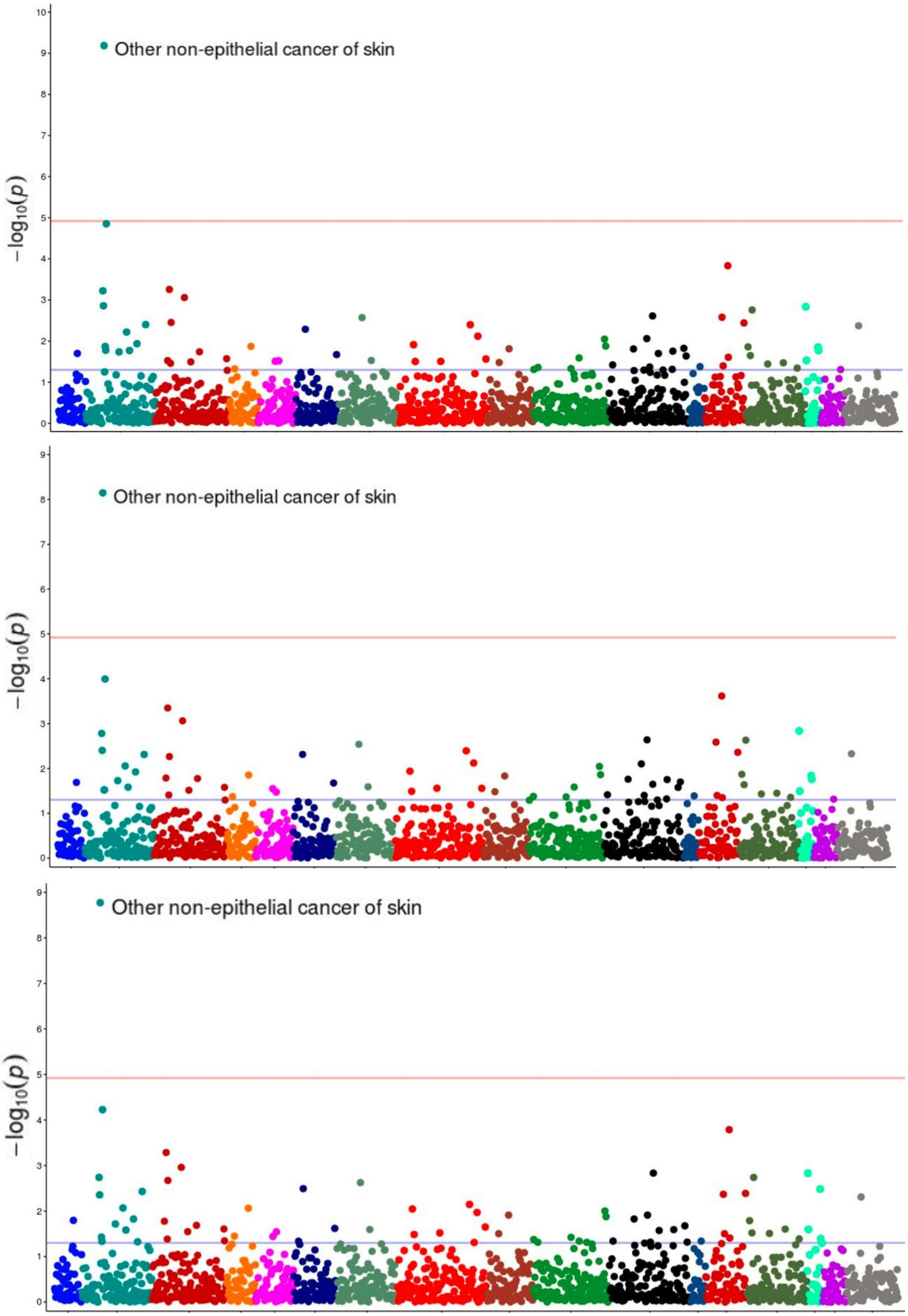


Figure 2.8: PheWAS analysis result for SNP rs1847134 with Gold-standard Logistic regression model (top), bias-corrected Poisson (middle) and robust Poisson model (bottom). Each dot represents one Phecode. The red line indicates the significance line. X axis denotes the negative log 10 p-values. The only selected code is Other non-epithelial cancer of skin. The maximum value of X axis is 10 for the top plot and 9 for the middle and bottom plot.")

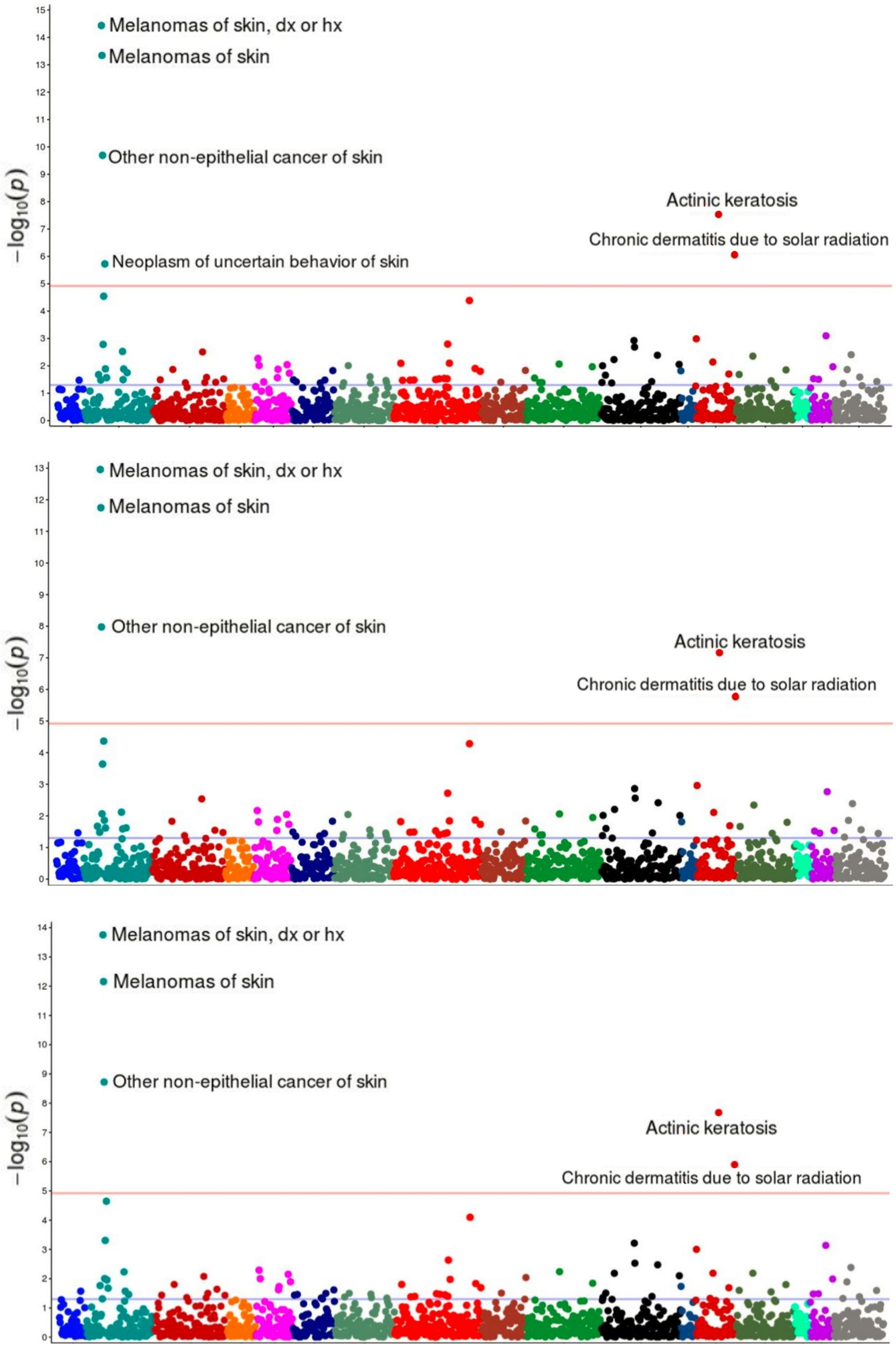


Figure 2.9: PheWAS analysis result for SNP rs258322 with Gold-standard Logistic regression model (top), bias-corrected Poisson (middle) and robust Poisson model (bottom). Each dot represents one Phecode. The red line indicates the significance line. X axis denotes the negative log 10 p-values. Top selected codes are Melanomas of skin, dx or hx, Other non-epithelial cancer of skin and Actinic Keratosis. The maximum value of X axis is 15 for the top plot; 13 for the middle and 14 for the bottom plot.")

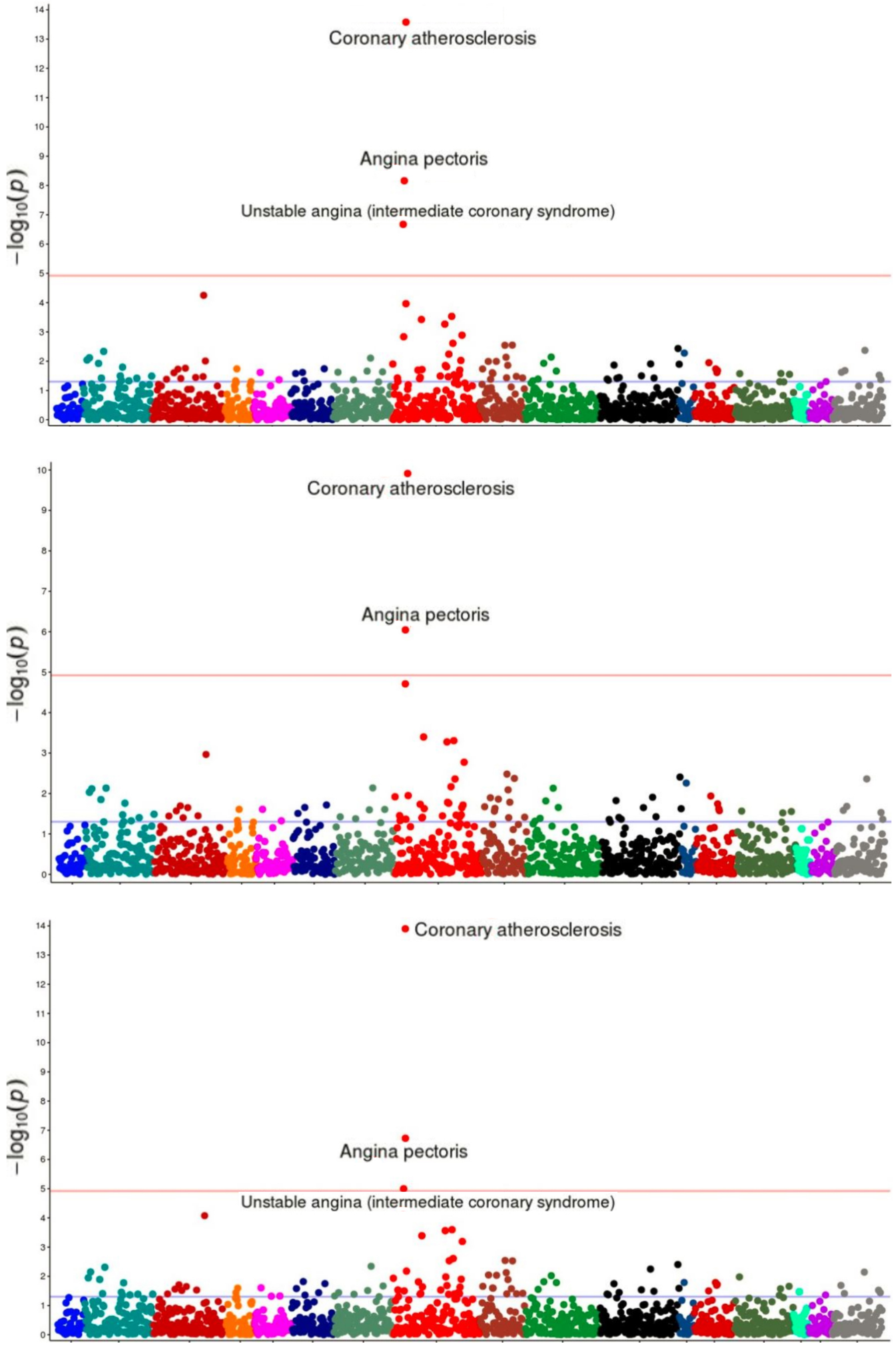


Figure 2.10: PheWAS analysis result for SNP rs4977574 with Gold-standard Logistic regression model (top), bias-corrected Poisson (middle) and robust Poisson model (bottom). Each dot represents one Phecode. The red line indicates the significance line. X axis denotes the negative log 10 p-values. Top selected codes are Coronary atherosclerosis, Angina pectoris and Unstable angina. The maximum value of X axis is 14 for the top plot; 10 for the middle and 14 for the bottom plot.")

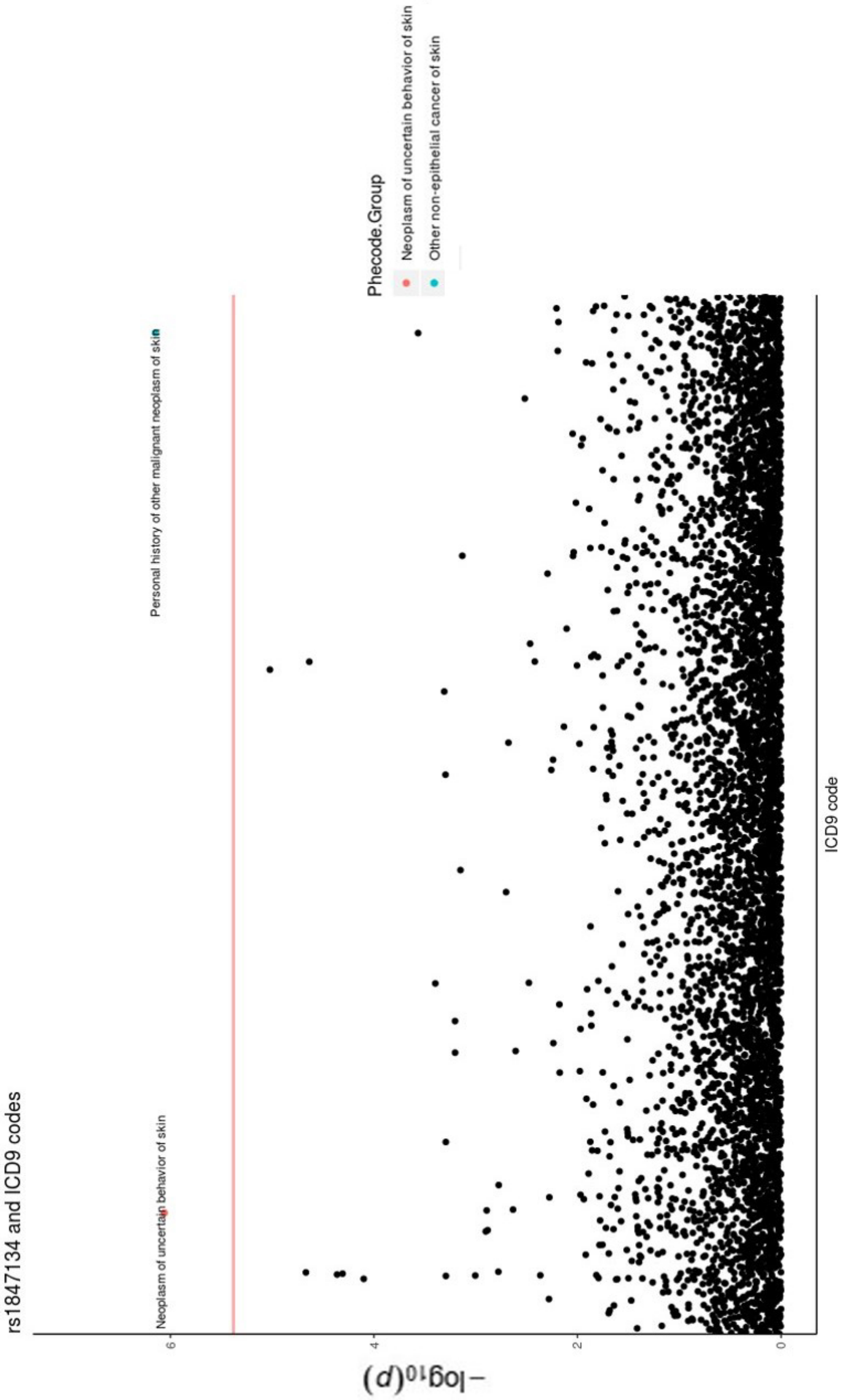


Figure 2.11: Result for ICD 9 analysis with SNP rs1847134. Each dot represents an ICD 9 code. Y-axis denotes the negative log 10 P-values. The colors of the dot corresponds to Phecode groups indicating in the legend. The red line indicates the significance line.")

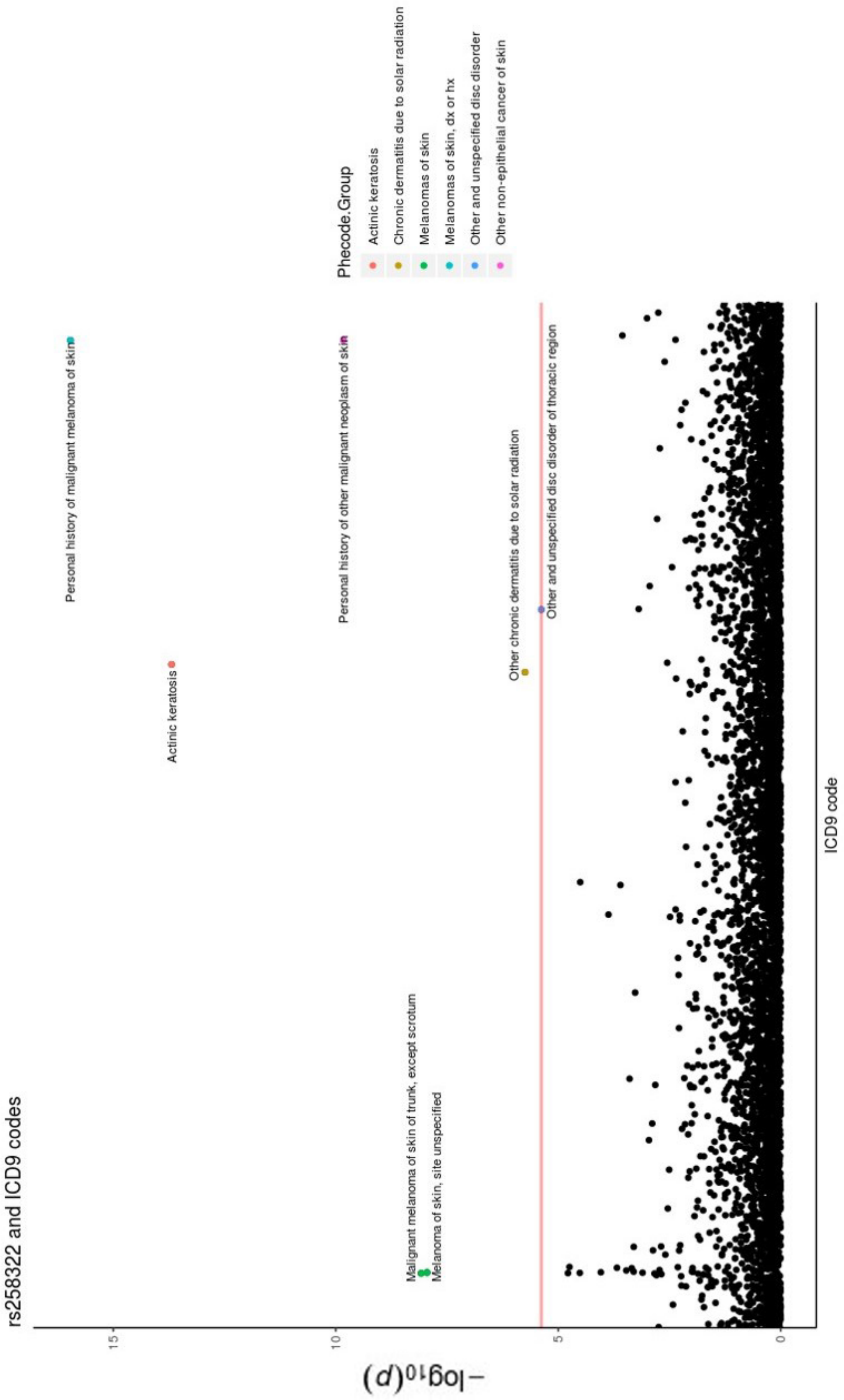


Figure 2.12: Result for ICD 9 analysis with SNP rs258322. Each dot represents an ICD 9 code. Y-axis denotes the negative log 10 P-values. The colors of the dot corresponds to Phecode groups indicating in the legend. The red line indicates the significance line.")

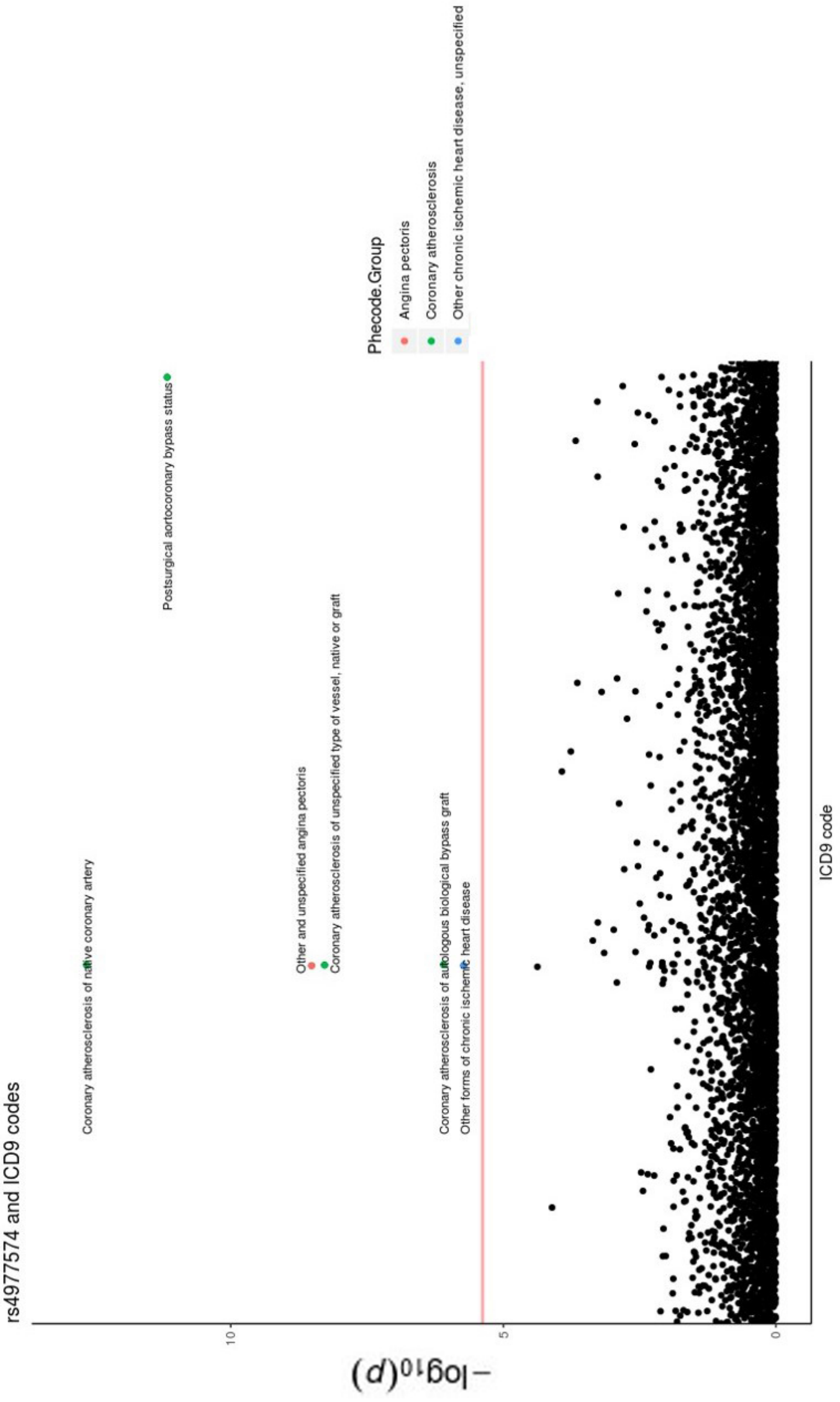


Figure 2.13: Result for ICD 9 analysis with SNP rs4977574. Each dot represents an ICD 9 code. Y-axis denotes the negative log 10 P-values. The colors of the dot corresponds to Phecode groups indicating in the legend. The red line indicates the significance line.")

CHAPTER 3

SUGGESTION FOR SPATIAL SMOOTHING IN FMRI GROUP INFERENCE : USE WITH CAUTION! – LIKELIHOOD SPATIAL SMOOTHING AS A MORE FLEXIBLE OPTION TO SMOOTH FMRI DATA

3.1 Introduction

Functional magnetic resonance imaging (fMRI) is one of the most widely used and a powerful tool to map brain activities by measuring changes in blood flow. A typical fMRI dataset consists of measuring the blood oxygen level-dependent (BOLD) contrast on three-dimensional volume elements, called voxels, over a set of discrete time series. (Ogawa et al. (1990); Ogawa et al. (1992); Kwong et al. (1992)). In general, fMRI data analysis can be conducted on voxel and region level. The voxel-wise approach measures the brain activity at the voxel levels while the region of interest (ROI) approach conducts analysis on groups of voxels at pre-defined regions by taking mean or median of the voxel-level measurements. Regardless of the types of analyses, ignoring either spatial or temporal correlation may lead to misleading conclusions. In fMRI data analysis, normal linear models are mainly used. As mentioned in Kang et al. (2012), most of the early work has been focusing on the modeling of temporal correlation in fMRI data. The commonly known work includes the autoregressive (AR(1)) models proposed by Bullmore et al. (1996) and general linear models applied to smoothed time series by Worsley and Friston (1995). Later, more work has been focusing on spatio-temporal correlation. One way to account for spatial correlation is spatial smoothing. It is usually incorporated as a part of the standard preprocessing steps for fMRI analysis together with scanner drift correction, motion correction, correction for cardiac and respiratory-related physiological noise, co-registration between the subject-specific anatomical and functional images, normalization. The biggest advantage of spatial smoothing step is to increase signal-to-noise ratio (SNR) for detection (Hopfinger et al. (2000); Bennett and Miller (2010); Mikl et al. (2008); Pajula and Tohka (2014)).

Worsley et al. (1996) addressed the spatial correlation by applying Gaussian kernel on the fMRI data. Due to the large usage of general linear model (GLM) in the fMRI analysis, Gaussian kernel spatial smoothing was widely accepted as part of standard preprocessing steps mainly for fulfilling the Gaussianity assumption of the GLM models (Mikl et al. (2008)). With the Gaussian kernel spatial smoothing, the observed value of the voxel in the center of the smoothing filter will be recomputed with the weighted averages of values from other voxels within the filter. The weights depend on the distance from the centric voxel and the joint Gaussian densities. Another type of spatial smoothing, proposed by Katanoda et al. (2002), accounts for the spatial dependency by borrowing information from the neighboring voxels in the Fourier domain. Instead of implementing the spatial smoothing on the original fMRI data as used in Gaussian kernel method, the method proposed by Katanoda et al. (2002) smoothed the likelihood functions instead. However, even though spatial smoothing has been accepted as a necessary step for preprocessing, little attention has been paid to the choice of smoothing methods or the degree of smoothing. The decisions made during the spatial smoothing step for fMRI group inference could potentially lead to inflation of false positives. Studies have shown that smoothing filter size, sample size and location of the brain regions (the inter-subject variability might be different between brain regions and leads to different SNRs) could be the potential contributing factors (Mikl et al. (2008); White et al. (2001)).

In this study, we first extended the method proposed by Katanoda et al. (2002) to time domain. The spatial smoothing was implemented on the regression coefficients from the GLM models with fMRI time series data. The final coefficient for a specific voxel was the weighted average of the coefficients from the neighboring voxels with inverse-variance weighting. We refer this method as "Neighboring voxel" method for the remaining of the text. Next, we evaluated the influence of degree of smoothing under combinations of different experimental settings, including sample size, SNR and length of time series for both Gaussian kernel method and Neighboring voxel method. To our knowledge, the influence of length of time series hasn't been assessed in previous studies before. Lastly, we briefly compare the performance of both spatial smoothing methods. An outline of the paper is as follows. In Section 3.2.1.1 and Section 3.3.1, we present simulation studies that illustrate the influence of degree of smoothing under different combination of experimental settings for both Gaussian kernel method and Neighboring voxel method. We further investigated how these factors could contribute to the potential inflation of false positives. In Section 3.2.1.2, 3.2.1.3 and 3.3.2, we apply the two smoothing methods to real fMRI data analysis. Finally, in Section 3.4, we summarize the results from the previous sections and provide some general advice for decision-making during spatial smoothing step.

3.2 Methods

3.2.1 Data

3.2.1.1 Simulated data

To simulate the spatio-temporal correlation similar to the real fMRI data, we adapted similar data generation process as described in Kang et al. (2015) where data were generated following the first order autoregressive model (AR(1)) in a region with spatial dimension 32 × 32 voxels. Data for multiple subjects were simulated instead of single subject as was in the original paper. Denote $Y_v(t)$ as the response at a voxel v (v = 1,..., V) and time t (t = 1,..., T) at a single subject level. The model with P stimuli can be expressed as the following for each subject:

$$Y_v(t) = \sum_{p=1}^{P} X_p(t)\beta_v^p + \epsilon_v(t)$$

where X_p denotes the convolution between the p^{th} stimulus impulse function and the hemodynamic response function (HRF). The canonical HRF function from Statistical Parametric Mapping 12 (SPM12) (Friston et al. (2007)) was used to generate $X_p(t)$. We assumed 2 boxcar stimuli for the simulation. The first stimulus was on during [1, $(T/4)$+1] and [$(T/2)$+1, $(3T/4)$+1] and the second is on otherwise. Spatial correlation was implemented on the data via exponential covariance function with the decaying parameter 2 and variance 2.5. The temporal correlation, $\epsilon_v(t)$ follows AR(1) process with AR(1) parameter 0.4 and standard deviation 1.5. We assumed that there are two active blocks with 64 voxels and 36 voxels. The data were either not spatially smoothed or smoothed by either Neighboring voxel method or Gaussian kernel method with varying filter sizes.

To compare results from different smoothing methods under various scenarios, we varied the number of time points, number of subjects and signal-to-noise ratios (SNRs). To be generalized to real fMRI data, we included T = 64, T = 128 and T = 256 for time points; 5, 10, 20 and 30 subjects and averaged SNRs: 0.087 (range: 0.0435 $\sim$ 0.131), 0.130 (range: 0.0870 $\sim$ 0.173), 0.18 (range: 0.135 $\sim$ 0.225), 0.26 (range: 0.173 $\sim$ 0.347) (calculated with the formula: effect size / standard deviation). The SNRs correspond to the reported effect sizes: 0,2, 0.3, 0.4 and 0.6. We included spatially unsmoothed, smoothed with Gaussian filter, FWHM = 2, 4, 8 mm and likelihood estimates smoothed with nearest neighbors and third-level neighbors. For all simulations, the null hypothesis of interest is H_o: $\beta^2 - \beta^1 = 0$ at each voxel. T-statistics and one-sided p-values were computed for all the smoothing methods compared. After adjusting for multiple comparison with FDR at 0.05, Type I error rate, Type II error rate and the mean error rate were used as comparison metrics.

3.2.1.2 Task-induced fMRI

Analysis on task-induced fMRI data was conducted to assess the influence of different spatial smoothing methods with varying degree of smoothing. The data were acquired from a sample of 29 right-handed postmenopausal women between the ages of 45 - 75 with major depressive disorder (MDD). The participants were asked to perform emotion dot probe (EDP) task, a spatial attention task that measures attention bias (Kimonis et al. (2006); Muñoz Centifanti et al. (2013)). The EDP task used in this study was a picture variant using images from the International Affective Pictures System (IAPS) (Lang et al. (1999)) and included neutral, positive, and negative (threat and distress) images. Trials of the EDP consisted of a fixation cross presented in the middle of the screen, followed by a brief presentation (500 ms) of a picture pair with one image each on the right and left of the screen. After the picture presentation, a target (asterisk) appeared either on the right or left of the screen (replacing one of the images) and the participant was instructed to indicate by finger button press the side of the screen on which the target appeared as quickly as possible. The EDP was adapted for fMRI and run as an event related design with three trial types relevant to the presented data: neutral-neutral pair, neutral-negative pair and neutral-positive pair. There were 5 stimuli measured in the study and we are interested in the contrast among two of the stimuli, measuring the brain activity difference between pressing on the button and seeing the asterisk when presented negative images.

Participants were scanned on a Philips 3.0 Tesla Achieva scanner, with eight channel head coil. Each subject received a sagittal T1-weighted 3D turbo field echo sensitivity encoding (TFE SENSE) sequence perpendicular to the anterior commissure (AC) -posterior commissure (PC) line, repetition time (TR) of 9.9 ms, echo time (TE) of 4.6 ms, a flip angle of 8 degrees, number signal averages (NSA) 1.0, a field of view (FOV) of 256 mm, a 256 x 256 matrix, and 1.0 mm slice thickness with no gap for 140 contiguous slices as well as blood oxygen level dependent (EpiBOLD) functional sequence during the EDP with transverse orientation, TR 2500 ms, TE 35 ms, flip angle 90 degrees, 1 NSA for, FOV 240, 240 x 128 matrix, and 4.0 mm slice thickness with no gap, with ascending interleaved acquisition, for 35 contiguous slices. The data were preprocessed using FSL and MATLAB version R2019 (MATLAB (2019)). The preprocessing steps included realignment of the functional runs and correction for bulk-head motion, co-registration of functional and anatomical images for each participant, segmentation of the anatomical image, normalization of the anatomical and functional images to the standard Montreal Neurological Institute (MNI) template. The preprocessed functional images had a voxel size of 2 × 2 × 2 mm. More details about the EDP task and imaging acquisition can be acquired in Albert et al. (2017). For Gaussian kernel methods, the preprocessed data were smoothed with Gaussian kernel filters at FWHM = 4 (Smooth.4), 6 (Smooth.6) and 8 (Smooth.8) mm. For the "No spatial" method, the preprocessed data were not spatially smoothed. For the Neighboring voxel methods, the spatial smoothing step was implemented on the regression coefficients from GLM models instead of the preprocessed data. Results smoothed with "Nearest neighbors" ($\sim$ 6 neighboring voxels in 3-D settings) and "Third-level neighbors" ($\sim$ 33 neighboring voxels in 3-D settings) were presented. The number of voxels involved in spatial smoothing is similar between Smooth.4 and Third-level neighbors. In the analysis, we included the 5 stimuli and 6 covariates derived from motion correction step. Denote the stimuli of interest, D1 and D2. The corresponding regression coefficients are β_v^{D1} and β_v^{D2}. The null hypothesis of interest is H_o: $\beta_v^{D1} - \beta_v^{D2} = 0$. T-statistics and 2 sided p-values were computed and controlled for multiple comparison at FDR = 0.05.

3.2.1.3 Resting-state fMRI

As a negative control for the inflation of Type I error rates observed in section 2, we evaluated the influence of smoothing methods by performing task-induced fMRI analysis on resting-state fMRI data, similar to what have been done previously in Eklund et al. (2016). By definition, the resting-state fMRI data are acquired when subjects are instructed to do nothing but lay still in the scanner without performing any specific task. The true effect for the stimuli should be 0. The data were acquired from a sample of 29 healthy volunteers between the ages of 20 and 50 years old. The subjects had no psychotropic medication use or history of psychiatric disorders. The patients were scanned on a Siemens 3.0 Tesla Trio Tim scanner with an eight channel head coil. Each subject received a T1 -weighted 3D magnetization-prepared rapid gradient-echo (MPRAGE) sequence with a repetition time of 2300 ms, echo time of 3.46 ms, a flip angle of 9 degrees with a voxel size of 0.9 × 0.9 × 1.2 mm as well as an EpiBOLD functional resting-state scan with repetition time of 2000 ms, echo time of 27 ms. The preprocessing steps include head motion correction across all scans, slice timing correction, co-registration and normalization to the standard MNI template. All preprocessing was performed using FSL software package (Smith et al. (2004)). The preprocessed functional images had a voxel size of 2 × 2 × 2 mm. For Gaussian kernel methods, the preprocessed data were smoothed with Gaussian kernel filters at FWHM = 6 (Smooth.6) and 8 (Smooth.8) mm. For the "No spatial" method, the preprocessed data were not spatially smoothed. For the Neighboring voxel method used here, the spatial smoothing step was implemented on the regression coefficients from GLM models by borrowing information from the nearest neighbors. The hypothesis testing steps generally followed the steps described at the end of subsection 3.1. Any detected signals were counted as false positives to compute Type I error rate. At sample size = 5, 10, 15 and 20, 300 random samples were selected for the analysis, e.g., for N = 10, 300 random combinations were selected out of $\binom{29}{10}$ combinations. The average Type I error rates were estimated and reported.

3.3 Results

3.3.1 Simulation study

We denoted "No spatial" for method without considering any underlying spatial correlation. Among the Neighboring voxel methods, results with "Nearest neighbors" ($\sim$ 4 neighboring voxels in 2-D simulation) and "Third-level neighbors" ($\sim$ 13 neighboring voxels in 2-D simulation) were presented. Among Gaussian kernel methods, results with FWHM = 2 ("Smooth.2"), 4 ("Smooth.4") and 8 ("Smooth.8") mm were included. A general rule of thumb for functional MR studies is that the Gaussian kernel filter size, FWHM, should be of the order of 2 to 3 times the voxel size (Worsley and Friston (1995); Newlander et al. (2014)). Therefore, FWHM = 2 mm was the minimum filter size selected in the simulation for Gaussian kernel method. With the formula, FWHM = 2.355 σ, the number of voxels involved in smoothing is similar between Smooth.2 and Third-level neighbors in the 2-D simulation. Lastly, we referred the term "larger power settings" to scenarios with larger sample size, longer time length or higher SNRs. The term "smaller power settings" was denoted otherwise. Type I, Type II and mean error rates between different smoothing methods for 5, 10, 20 and 30 subjects with T = 64, 128 and 256 and varying SNRs are presented in Figure 3.1, Figure 3.2, Figure 3.7 (Appendix) and Figure 3.8 (Appendix). Within each fixed sample size, Type II error rates decrease but Type I error rates increase as the length of time points and SNRs increase for all methods. In general, we observe that Type I error rates increase and Type II error rates decrease as the degree of smoothing increases (Gaussian kernel methods with larger filter sizes or Neighboring voxel methods with more neighboring voxels involved).

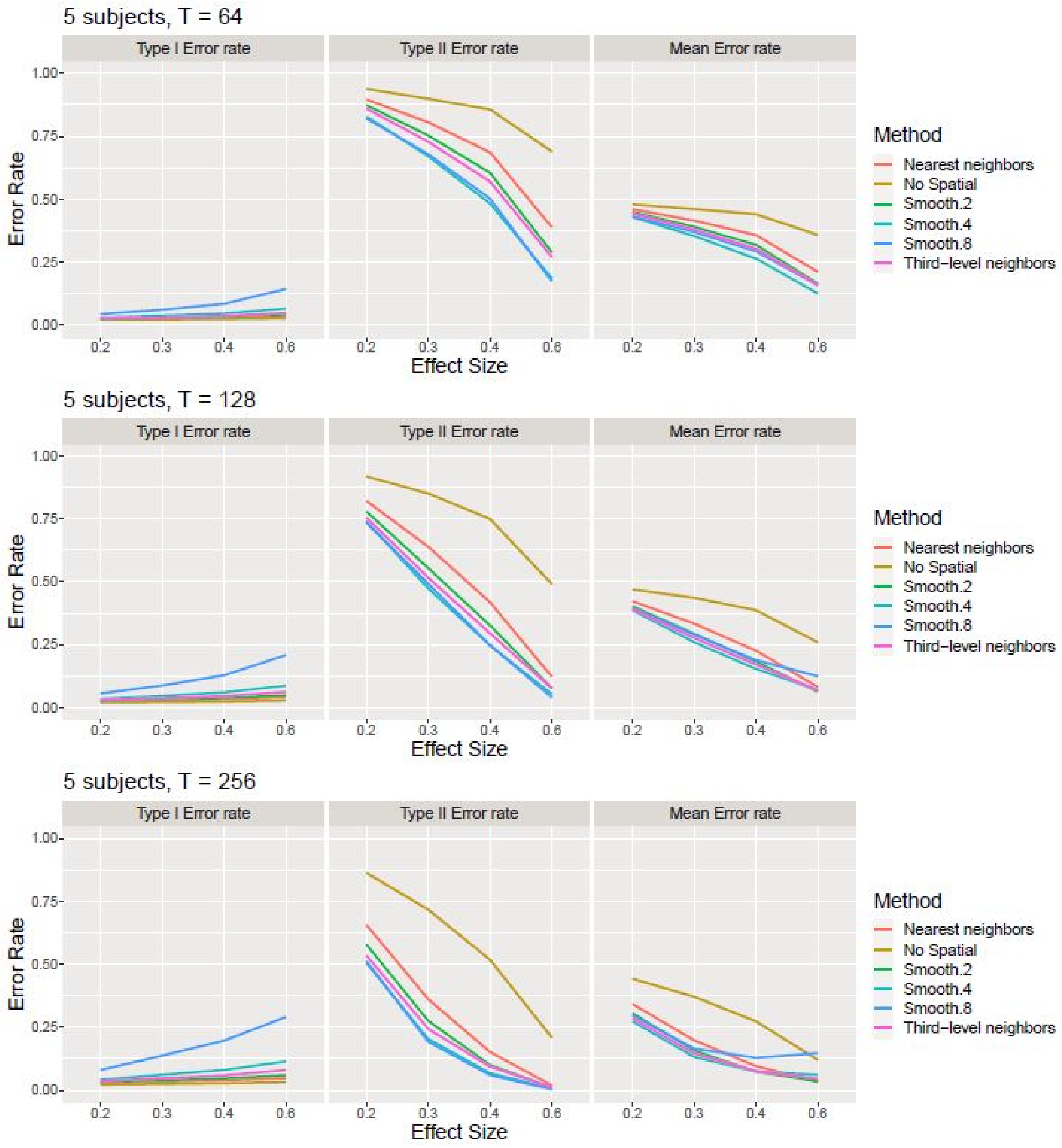


Figure 3.1: Simulation results for 5 subjects with different time lengths, T = 64, 128, 256. X-axis consists of the effect sizes, 0.2, 0.3, 0.4 and 0.6. The first column corresponds to the Type I error rates (average number of false positives divided by the total number of null voxels). The second column corresponds to the Type II error rates (average number of false negatives divided by the total number of non-null voxels). The third column corresponds to the mean error rates (average between Type I error rates and Type II error rates). No Spatial corresponds to method without considering underlying spatial correlation. Smooth 2, 4, 8 corresponds to Gaussian Kernel Filter sizes, FWHM = 2, 4, 8 mm. Nearest and Third-level neighbors correspond to Neighboring voxel method involving first-level closest and third-level closest neighboring voxels.

Next, we look into the performance of each method more closely in this simulation study. Without considering the underlying spatial correlation, the No spatial method

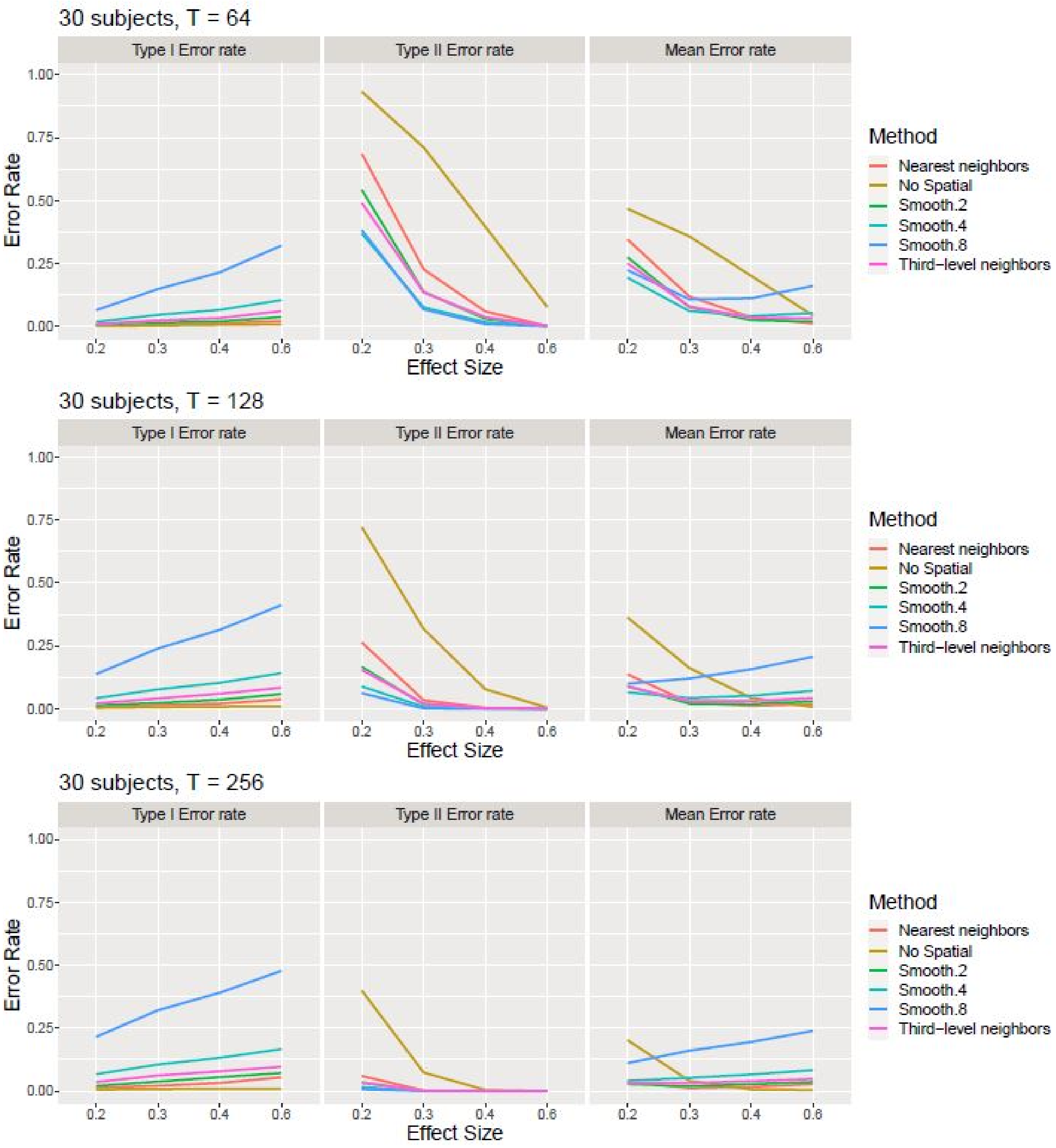


Figure 3.2: Simulation results for 30 subjects with different time lengths, T = 64, 128, 256. X-axis consists of the effect sizes, 0.2, 0.3, 0.4 and 0.6. The first column corresponds to the Type I error rates (average number of false positives divided by the total number of null voxels). The second column corresponds to the Type II error rates (average number of false negatives divided by the total number of non-null voxels). The third column corresponds to the mean error rates (average between Type I error rates and Type II error rates). No Spatial corresponds to method without considering underlying spatial correlation. Smooth 2, 4, 8 corresponds to Gaussian Kernel Filter sizes, FWHM = 2, 4, 8 mm. Nearest and Third-level neighbors correspond to Neighboring voxel method involving first-level closest and third-level closest neighboring voxels.

is the most conservative smoothing method throughout all scenarios. The trade-off of higher Type II error rates for lower Type I error rates leads to the largest mean error rates in most settings. Among Gaussian kernel methods, Smooth.4 performs the best in terms of mean error rate in lower power settings especially when N = 5 (Figure 3.1) and N = 10 (Appendix, Figure 3.7). In contrast, Smooth.2 performs the worst in terms of mean error rate in most lower power settings by having larger Type II error rates. As the power of the settings increases, the performance of Smooth.8 becomes the worst by having larger Type I error rates. It's worth mentioning that as the power of the settings increases, Smooth.2 starts to outperform Smooth.4. The observation is more obvious when N = 20 (Appendix, Figure 3.8) and N = 30 (Figure 3.2). Among Neighboring voxel methods, Third-level neighbors method outperforms the Nearest neighbors method by having lower Type II error rates and similar Type I error rates in lower power settings. In larger power settings, Nearest neighbors method has lower mean error rates with the advantage of having lower Type I error rates. Similar to Gaussian kernel methods, Nearest neighbors method starts to outperform Third-level neighbors method as the power of the settings increases. This implies the potential need of smaller filter size/number of neighboring voxels for smoothing methods in higher power studies.

According to the above results, Type II error rates naturally decrease for all methods as the power of the studies increases. Larger degree of spatial smoothing decreases Type II error rates more. Nevertheless, the decreasing Type II error rates benefits come with the price of increasing Type I error rates. We further explore the reasoning behind the false positives and how experimental factors relate to the false positives. We presented the average proportion of being classified as significant for each voxel at N = 20 with T= 128 (Figure 3.3) and T= 256 (Appendix, Figure 3.9). The voxels within the red boxes are the true active voxels. We included the results from No spatial method, Third-level neighbors method, Smooth.4 and Smooth.8 (increasing order of degree smoothing). In both Figure 3.3 and Figure 3.9 (Appendix), we observe that only parts of the voxels within the red boxes are classified as active with 50% of less for No spatial method. As the degree of smoothing increases, the true active voxels are detected more often. In addition, the voxels surrounding the true active voxels are classified as significant more frequently as the degree of smoothing increases. These voxels actually contribute to Type I error rates. Comparing to Figure 3.3, we notice that as the length of time series increases from 128 to 256 (Appendix, Figure 3.9), more voxels surrounding the true active voxels were misclassified as significant. The observation in Figure 3.3 and Figure 3.9 (Appendix) implies that the false positives occur near the true active voxels. As the power of the studies increases, more voxels adjoining the true active voxels are misclassified as active voxels.

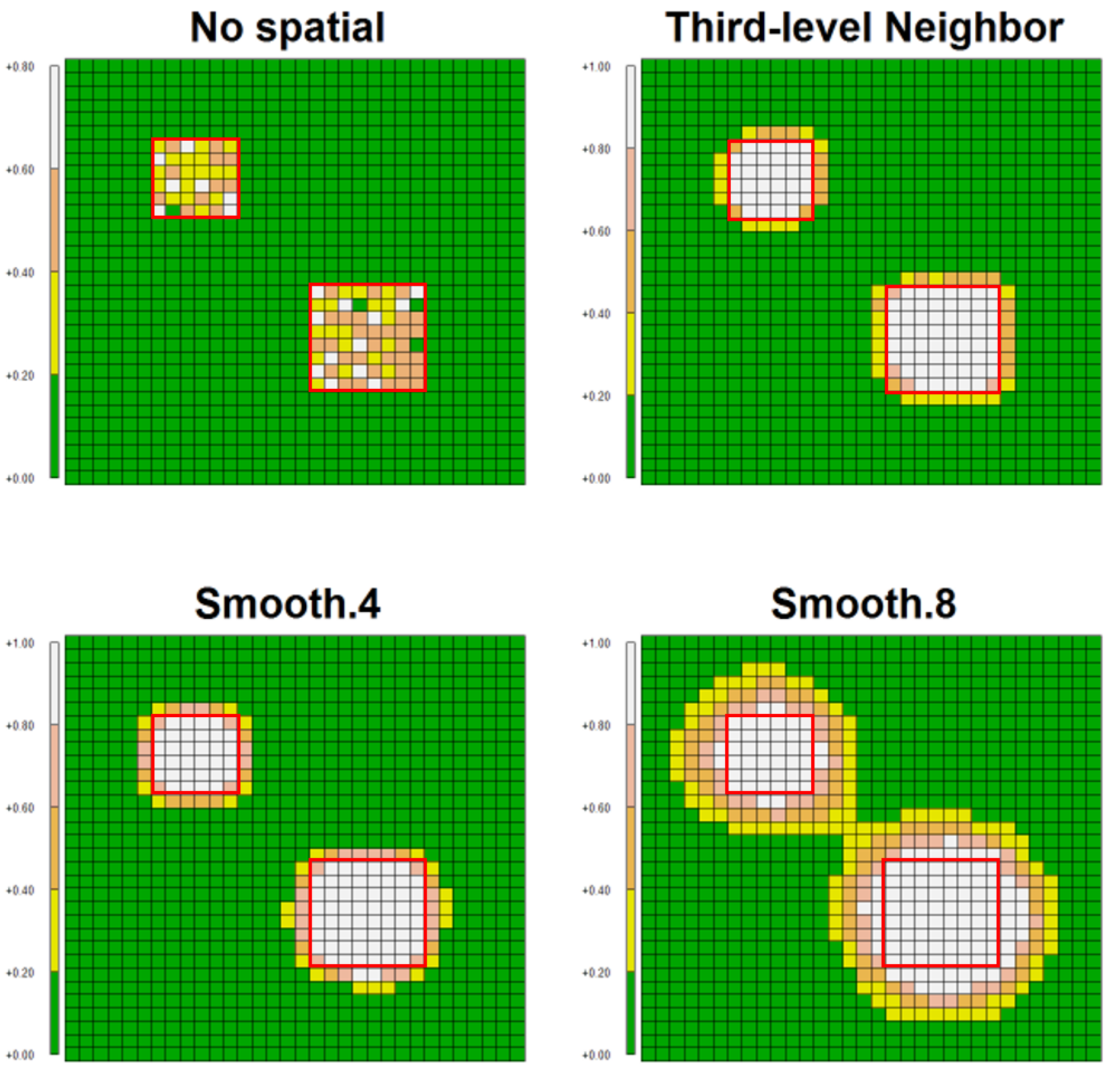


Figure 3.3: Average proportion of being classified as significant for each voxel over 300, 32 by 32, 2D simulations with effect size = 0.3, sample size = 20 and T = 128. Results from No spatial, Third-level neighbors, Smooth.4 and Smooth.8 are reported. The color bar indicates the average proportion of a voxel being classified as significant. The bars are in the increments of 0.2, ranging from 0 to 0.8 for No spatial method and 0 to 1 for the others. The voxels within the red boxes are truly active.

The direct comparison between the performance of Gaussian kernel methods and Neighboring voxel methods could be difficult as the number of voxels involved in smoothing and how these voxels contribute to smoothing are different. With the formula, FWHM = 2.355 σ, the number of voxels involved in smoothing is approximately the same between Smooth.2 and Third-level neighbors method. Here, we briefly compare the performance between these two methods. When T = 64 in Figure 3.1, Smooth.2 has slightly lower Type I error rates and higher Type II error rates compared to Third-level neighbors method. In all other scenarios, Smooth.2 has similar Type II error rates while having a bit lower Type I error rates compared to Third-level neighbors method, leading to somewhat better performance in terms of mean error rates for Smooth.2 in the simulation. In higher power settings, especially when subject sample size is at least 20 and length of time series is at least 128, the Nearest neighbors method starts to outperform Smooth.2 by having lower Type II Error rates. Since FWHM = 2 mm is the lowest filter size for Gaussian kernel methods, this indicates the need to go beyond the minimum filter size for Gaussian kernel methods in certain scenarios and that Neighboring voxel methods might be more flexible in adjusting the degree of smoothing.

3.3.2 Clinical application

We first performed task-induced fMRI analysis on resting-state fMRI data and the result is shown in Figure 3.4. We notice that the Type I error rate does not depend on the smoothing degree. In addition, the Type I error rate drops to 0 as sample size increases. This result demonstrates opposite trend when performing analysis on task-induced fMRI data. It also serves as a negative control for the Type I error rate when the true effect size is 0. When the true effect size is 0, the Type I error rate does not depend on smoothing degree and decreases as sample size increases.

The analysis results for task-induced fMRI data are shown in Figure 3.5, Figure 3.6 and Figure 3.10 (Appendix). In the simulation section, we observed that smoothing degree as well as experimental factors both affect activation. In Figure 3.10 (Appendix), when sample size = 15, only one side of the middle temporal gyrus is classified as active while in full sample size, both sides of middle temporal gyrus are classified as active. This potentially indicates that the power increases as sample size increases. For the occipital lobe area, we observe that the activation areas expand as sample size increases to 29. In Figure 3.5, activation maps of varying smoothing degree are shown. Both Gaussian kernel and Neighboring voxel methods follow the trend that as the filter size/number of neighboring voxels increase, the activation areas expand. No spatial model is the most conservative method as observed in the simulation section. In Figure 3.5A, although sparse, most of the activation surrounds the cuneus region followed by some activation in lingual gyrus. These regions are located in the occipital lobe. In Figure 3.5B, by incorporating the nearest neighbors, stronger signals are seen in both cuneus and lingual gyrus regions. With

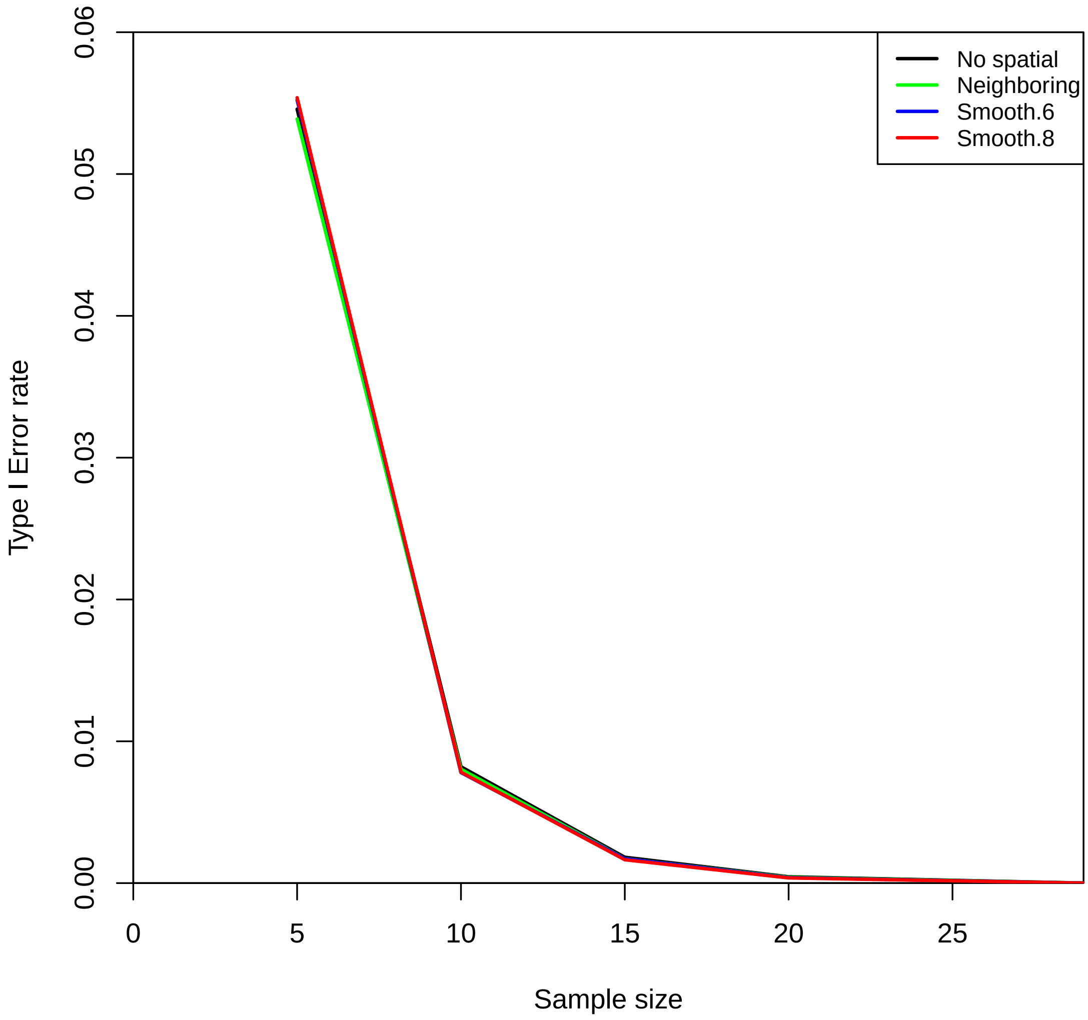


Figure 3.4: Relationship between average Type I error rate and sample size for No spatial, Smooth.6, Smooth.8 and Nearest neighbors method after implementing whole-brain task-induced fMRI analysis on resting-state fMRI data. 300 resamples from full sample size at 29 were drawn with sample size 5, 10, 15 and 20 to compute the average Type I error rate.

third-level neighbors incorporated (Figure 3.5C), more activation is observed in occipital lobe and more obvious activation is shown on both sides of cuneus and lingual gyrus regions. In Gaussian kernel methods, we observe similar activation patterns as filter size increases. In real data, we can roughly approximate the Third-level neighbors method with Smooth.4 as shown in Figure 3.6. Although showing slight increase of Type I error rates in the simulation results, Neighboring voxel method seems to perform similarly with Gaussian kernel method in real data. So far, we have

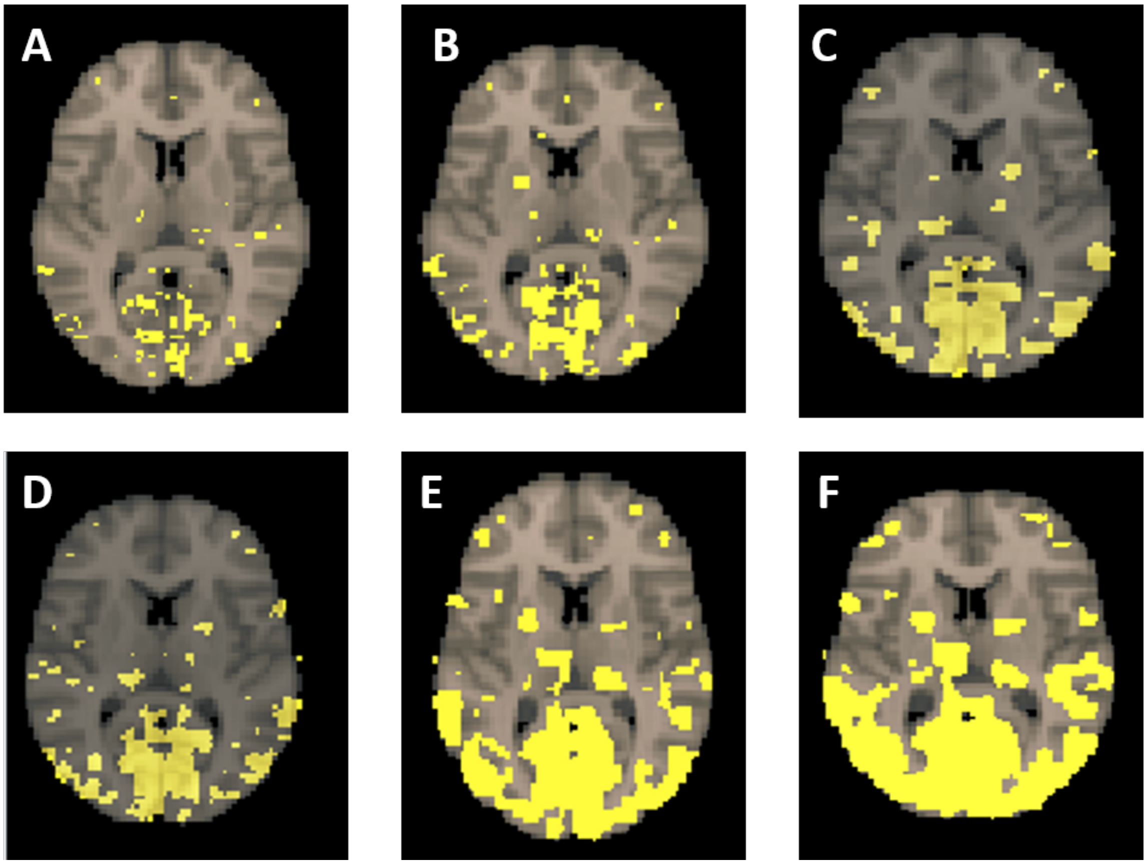


Figure 3.5: Activation maps for the 37th axial slices when sample size = 29. The yellow blobs indicate activation areas. A corresponds to No Spatial method; B corresponds to Nearest neighbors method; C corresponds to Third-level neighbors method; D corresponds to Smooth.4 method; E corresponds to Smooth.6 method and F corresponds to Smooth.8 method.

demonstrated that in real data, both experimental settings and smoothing degree are important for the classification of active regions and highly related to Type I and II error rates. The performance of Gaussian kernel method and Neighboring voxel method are approximately identical.

We further notice that activation in different regions are heterogeneous. In Figure 3.5, the true activation might be mainly located in the occipital lobe, cuneus, lingual gyrus and middle frontal gyrus. However, the activation of middle frontal regions are only noticeable in Smooth.6 (Figure 3.5E) and Smooth.8 (Figure 3.5D) while excess activation is observed in the bottom half of the brain. This indicates that the SNRs are different between each active region. This observation is consistent with previous literature that some ROIs, if are truly active, would require smaller filter sizes while others require larger filter sizes (White et al. (2001); Mikl et al. (2008)). In summary, the choice of the smoothing degree needs to be carefully selected and that uni-filter

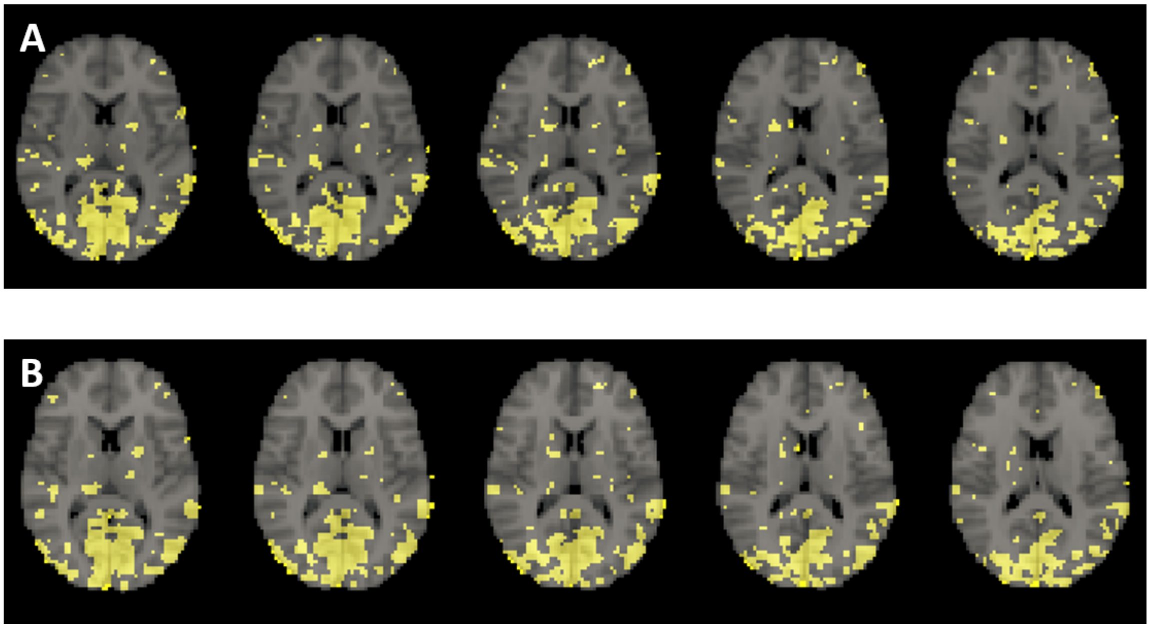


Figure 3.6: Activation maps for the 41th to 45th axial slices when sample size = 29. The yellow blobs indicate activation areas. A corresponds to Smooth.4 method and B corresponds to Third-level neighbors method.

size/number of neighboring voxels might not be appropriate for smoothing procedure in whole brain analysis.

3.4 Discussion

Inflation of Type I error rates due to spatial smoothing has been observed in our study and reported in some previous literature. Our study further concludes that the falsely active voxels mainly surround the active voxels. The degree inflation of Type I error rates is influenced by the smoothing degree and experimental settings. Increasing SNRs, sample size and length of time series lead to an increase in power and Type I error rates. The amount of the inflation depends on the smoothing degree. Especially in Smooth.4 and Smooth.8, the increase in Type I error rates is more profound compared to methods with smaller smoothing degree.

Spatial smoothing is a standard and needed procedure for current fMRI analysis. However, the choice of specific filter size is often arbitrary. Results have indicated that careful consideration on the choice of degree of smoothing is required to obtain valid inference and that both experimental settings and ROIs (due to heterogeneous activation between different regions) are important factors when making decisions on spatial smoothing. Based on the results shown in this study so far, we can infer that one size of filter/fixed number of neighboring voxels for whole brain analysis might not be appropriate as mentioned in subsection 3.3. Some research have been focusing on determining the optimal Gaussian kernel filters sizes (Mikl et al. (2008); White et al. (2001)). Some has focused on using different kernel shapes other than Gaussian kernel (Bartés-Serrallonga et al. (2014)). Lastly, others have focused on adaptive Gaussian kernel smoothing on images from a single time series (Yue et al. (2010); Strappini et al. (2016)). Specifically, Yue et al. (2010) points out that there are no guarantees of improved inference on group level with the adaptive smoothing approaches and it is currently not feasible to optimally smooth all images simultaneously. On a different direction, Wang et al. (2013) focused on estimating the accurate HRF instead and has shown promising results. However, the computation can be burdensome and the selection of optimal bandwidth still requires further research. With the scale of the whole brain voxel-wise analysis, further research in optimal degree of smoothing under different experimental settings and different regions of the brain seems to still be the most promising solution. In our study, we compared the behavior of two smoothing methods, Gaussian kernel method and Neighboring voxel method under different scenarios. Although these two methods performed similarly, Neighboring voxel method possesses some advantages over the traditional Gaussian kernel method. The number of neighboring voxels and the shapes of the smoothing filters are more flexible. In addition, edge effect is easier to be avoided with Neighboring voxel method.

In this study, we thoroughly explored the behavior of error rates and consequences from inappropriate spatial smoothing. Based on current group inference framework, research investigating the optimal spatial smoothing filter sizes considering regions of the brain and experimental settings might be the best solution to balance between Type I and II error rates. Currently, few research have been done in this area and little filter size advice can be provided for future analysis. We also evaluated an alternative smoothing method and pointed out that although Neighboring voxel method performs similarly to Gaussian kernel method, it provides several advantages over Gaussian kernel method as a tool for further research in optimal spatial smoothing. The research results can then be provided as suggestions for future fMRI analysis to obtain valid inference with appropriate spatial smoothing. Lastly, we emphasize again the careful consideration of the utilities for balancing between Type I and II error rates when applying spatial smoothing in group inference task-induced fMRI analysis.

3.5 Appendix

3.5.1 Neighboring voxel method

The Gaussian kernel smoothing method is widely used and can be directly implemented from MATLAB (MATLAB (2019)). Here we smooth the likelihood functions with neighboring voxels in time domain. There are total of V voxels, P external stimuli, total time points T and J subjects in the analysis. We define the time series at voxel v for subject j to be $Y_{j,v}(t)$ where t = *1,...,T*, j = *1,...,J* and v = *1,...,V*. The GLM model for each voxel v, subject j and time t looks like the following:

$$Y_{j,v}(t) = \sum_{p=1}^{P} \beta^p_{j,v} X_{j,p}(t) + \epsilon_{\text{j,v}}(t)$$

where

- $\beta^p_{j,v}$ is the coefficient estimate for voxel v due to stimulus p for subject j.
- $\epsilon_{j,v}(t)$ is the error term that accounts for temporal correlation under the additivity and separability assumption of spatio-temporal correlation for subject j. For temporal elements, we will assume that ϵ_v is independent of $\epsilon_{v'}$.

3.5.1.1 Overview of estimation

There are 2 stages for the estimation procedure. For stage 1, We obtained $\beta^p_{j,v}$ with iterative weighted least square (IWLS) method and retained its covariance matrix, Cov($R_{j,v}$). Here we provided steps to estimate these two parameters.

- Step 1: Obtain OLS estimate of $\widetilde{\beta^p_{j,v}} = (X^T_{j,p} X_{j,p})^{-1}(X^T_{j,p} Y_{j,v})$
- Step 2: Obtain residuals R_j where $R_{j,v} = Y_{j,v}$ - $\widetilde{\beta^p_{j,v}}$
- Step 3: Estimate $\beta^p_{j,v}$ and Cov($R_{j,v}$) with the following sub-steps
  - Step a: Assuming AR(1) parametric form, we can obtain $\hat{\phi}$ and $\hat{\sigma}^2$ with Yule-Walker equation to construct Cov($R_{j,v}$). The AR(1) formula can be written as the following linear regression form:
    $(R_{j,v,2}, ..., R_{j,v,T})^T = (R_{j,v,1}, ..., R_{j,v,(T\text{-}1)})^T \phi + (\varepsilon_{\text{j,v,2}}, ..., \varepsilon_{\text{j,v,T}})^T$
    We can obtain $\hat{\phi}$ and $\hat{\sigma}^2$ with OLS estimation.
  - Step b: Update and iterate Step 1 -3 until convergence of $\beta^p_{j,v}$. We can construct $\beta_{j,v}$ as vertical stack of $\beta^p_{j,v}$ for all stimuli. $\beta^p_{j,v} = (X^T \widetilde{\Omega}^{-1} X)^{-1}(X^T \widetilde{\Omega}^{-1} Y)$ where $\widetilde{\Omega}^{-1}$ is Cov($R_{j,v}$)

For stage 2, once we obtained the converged $\beta_{j,v}$ and its covariance matrix $\mathrm{Cov}(R_{j,v})$, we were able to start the smoothing process. We first defined the neighboring voxels to include. The neighboring voxels were determined based on the Euclidean distance between voxels in our settings. We included the nearest L neighbors with a total of *L + 1* voxels. The smoothing procedure has the formula, $\beta_j = F\beta_{s,j} + \epsilon'_{\mathrm{j}}$ where β_j is a vector with vertical stack estimates ($\beta_{j,v}$) from stage 1 for *L + 1* voxels (Dimension: $[P \times (L+1)] \times 1$). $\beta_{s,j}$ (Dimension: P × 1) is a vector with vertical stack of final smoothed estimates for P stimuli and subject j taking into account of L neighbors. F is a transition matrix with vertical stack of *L + 1* identity matrices with dimension of $P \times P$ (F dimension:$[P \times (L+1)] \times P$) . The covariance matrix for β_j is the block diagonal matrix of $\mathrm{Cov}(R_{j,v})$ for all *L + 1* voxels. With these elements, one can estimate $\beta_{s,j}$ with inverse-variance weighting method. For each voxel, $\beta^p_{s,j}$ denotes the smoothed estimate for subject *j*, stimulus *p* accounting for nearest *L* neighbors.

3.5.1.2 Group inference

The goal for the analysis is to distinguish active and inactive voxels when certain stimuli are presented to a group of subjects. We formed the hypothesis with the linear combination of β_s, that is $H_o : \beta^p_s - \beta^1_s = 0$ vs. $H_1 : \beta^p_s - \beta^1_s > 0$ where β_s are the group-level parameter of interest. To obtain group level inference, $\beta^p_{s,j} - \beta^1_{s,j}$ was obtained for each subject from stage 2 estimation. The mean and standard error of all subject-level linear combination were used to compute t-statistics for the hypothesis testing. The one-sided hypothesis was tested at voxel-level correcting for multiple comparison.

3.5.2 Figures

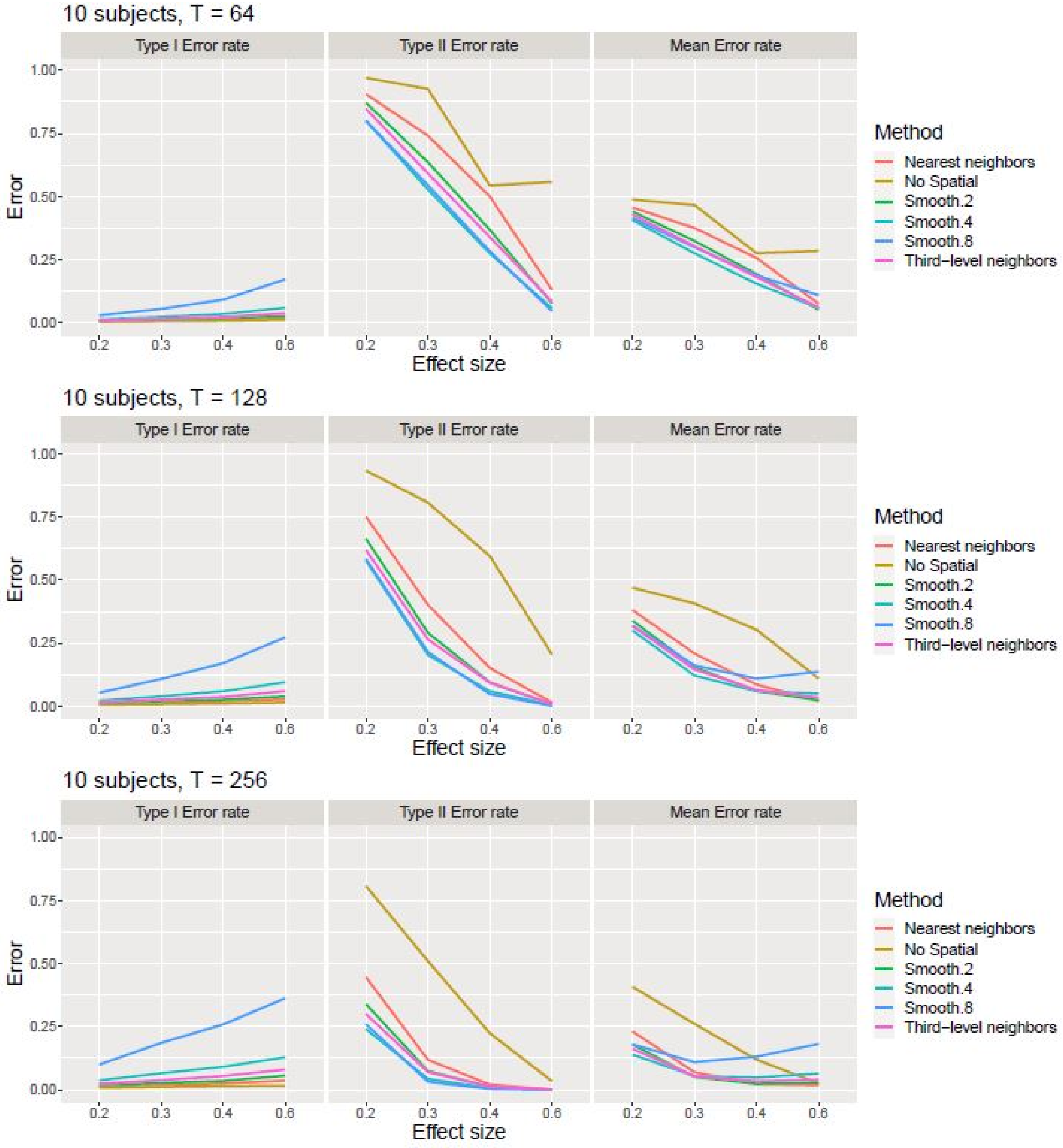


Figure 3.7: Simulation results for 10 subjects with different time lengths, T = 64, 128, 256. X-axis consists of the effect sizes, 0.2, 0.3, 0.4 and 0.6. The first column corresponds to the Type I error rates (average number of false positives divided by the total number of null voxels). The second column corresponds to the Type II error rates (average number of false negatives divided by the total number of non-null voxels). The third column corresponds to the mean error rates (average between Type I error rates and Type II error rates). No Spatial corresponds to method without considering underlying spatial correlation. Smooth 2, 4, 8 corresponds to Gaussian Kernel Filter sizes, FWHM = 2, 4, 8 mm. Nearest and Third-level neighbors correspond to Neighboring voxel method involving first-level closest and third-level closest neighboring voxels.

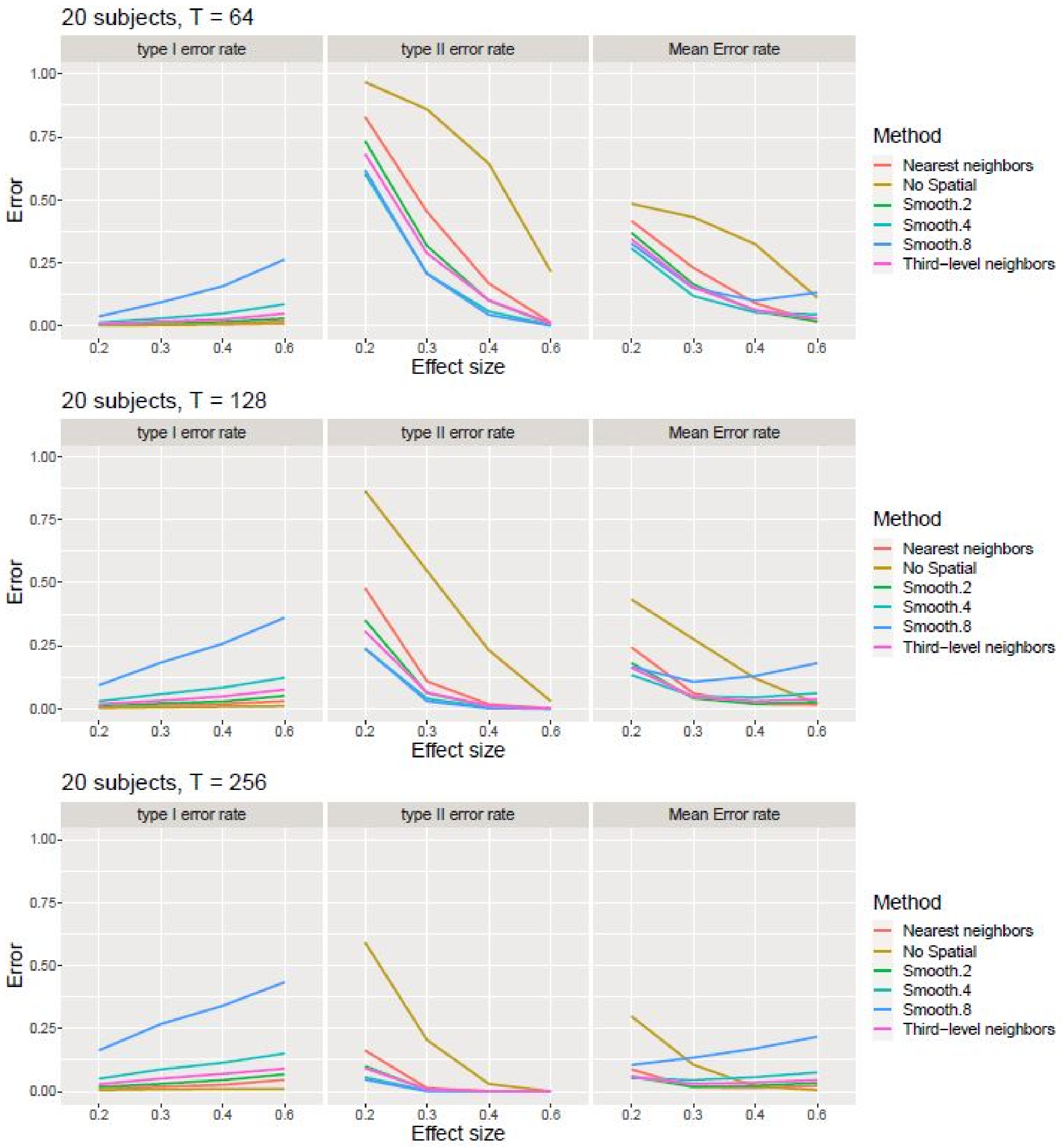


Figure 3.8: Simulation results for 20 subjects with different time lengths, T = 64, 128, 256. X-axis consists of the effect sizes, 0.2, 0.3, 0.4 and 0.6. The first column corresponds to the Type I error rates (average number of false positives divided by the total number of null voxels). The second column corresponds to the Type II error rates (average number of false negatives divided by the total number of non-null voxels). The third column corresponds to the mean error rates (average between Type I error rates and Type II error rates). No Spatial corresponds to method without considering underlying spatial correlation. Smooth 2, 4, 8 corresponds to Gaussian Kernel Filter sizes, FWHM = 2, 4, 8 mm. Nearest and Third-level neighbors correspond to Neighboring voxel method involving first-level closest and third-level closest neighboring voxels.

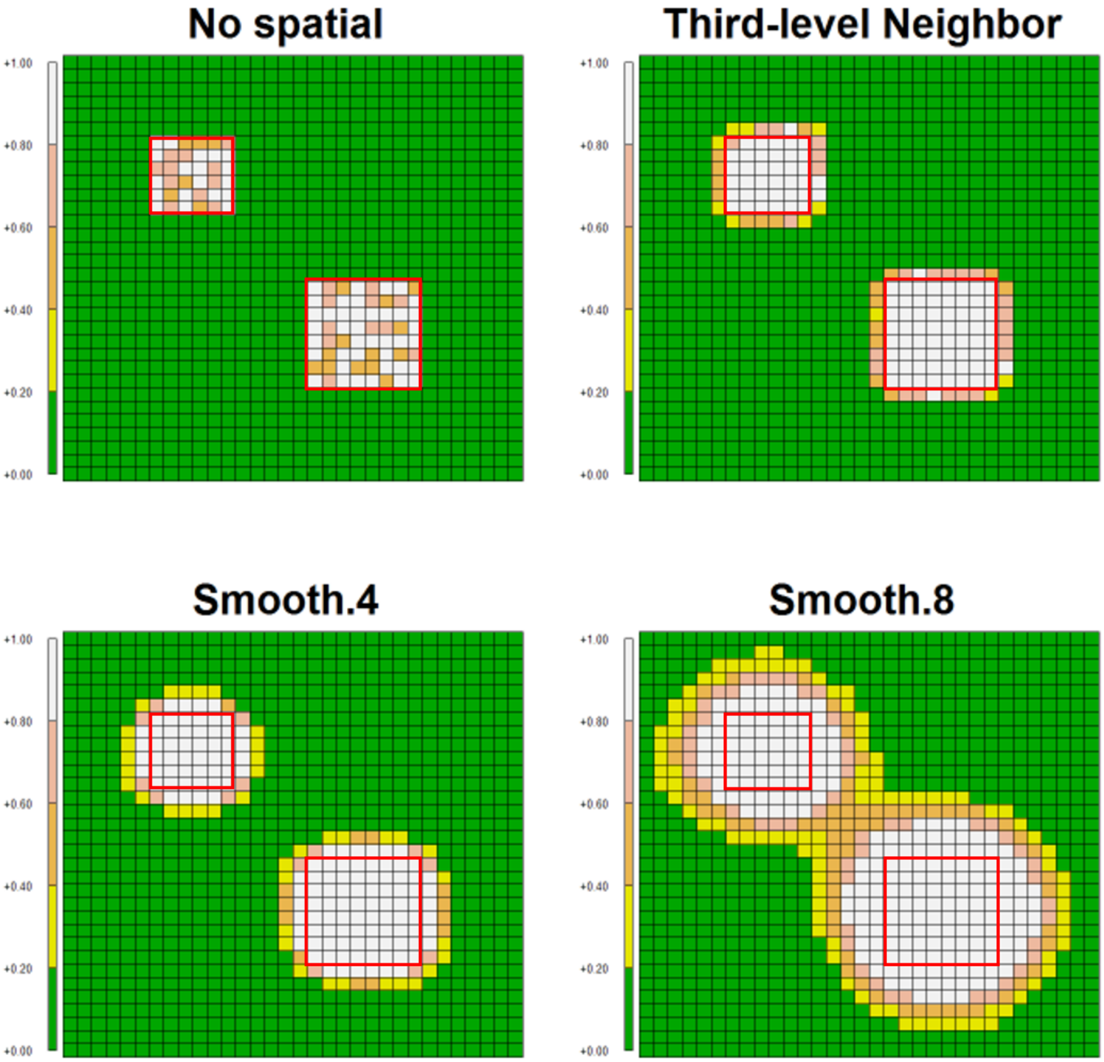


Figure 3.9: Average proportion of being classified as significant for each voxel over 300, 32 by 32, 2D simulations with effect size = 0.3, sample size = 20 and T = 256. Results from No spatial, Third-level neighbors, Smooth.4 and Smooth.8 are reported. The color bar indicates the average proportion of a voxel being classified as significant. The bars are in the increments of 0.2 from 0 to 1. The voxels within the red boxes are truly active.

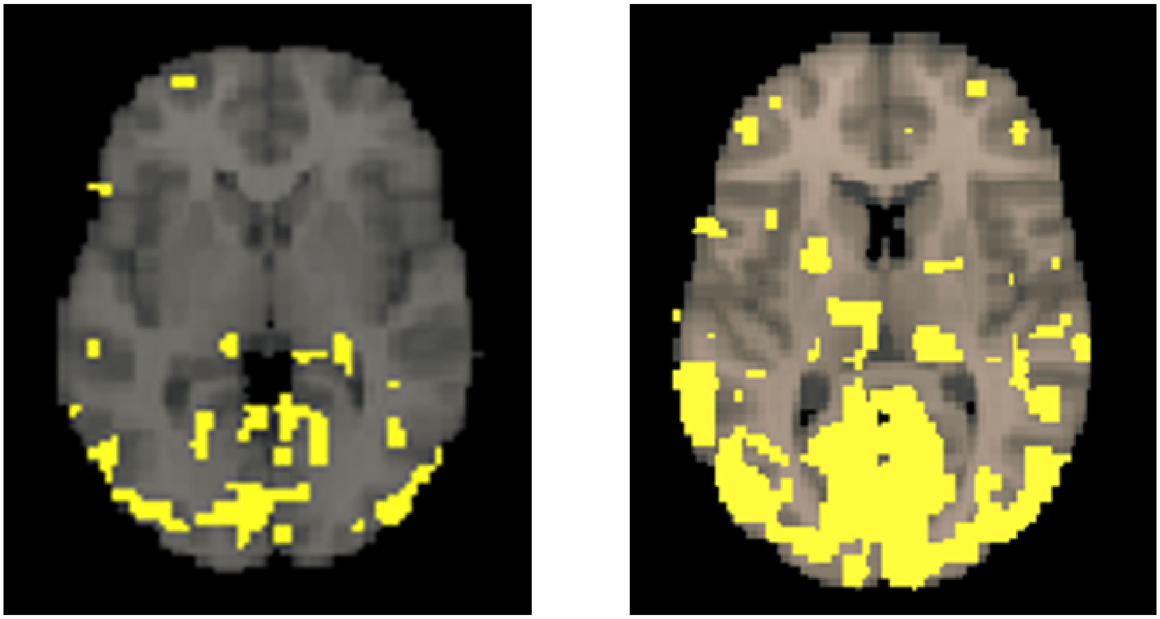


Figure 3.10: Activation maps for the 37th axial slice where the yellow blobs indicate activation areas resulting from Gaussian kernel method with FWHM = 6 mm. The left plot corresponds to sample size = 15. The right plot corresponds to full sample size = 29 subjects.

CHAPTER 4

SECOND-GENERATION P-VALUES FOR FUNCTIONAL MAGNETIC RESONANCE IMAGING DATA

4.1 Introduction

General linear model (GLM) followed by multiple testing correction on the reported p-values has remained the main hypothesis testing framework in functional magnetic resonance imaging (fMRI) data analysis. Due to the nature of fMRI data, large noises presented in the data require higher trade-off of Type I error rate for power in order to make meaningful scientific discoveries. Further, large number of comparisons on the voxel level analysis requires proper multiple comparison adjustment to draw appropriate inference. Lindquist and Mejia (2015) provided a coherent and detailed introduction to different multiple comparison methods used in the Neuroimaging field. There are two major types of multiple comparison adjustment methods commonly used in fMRI analysis. Random Field Theory (RFT) was proposed by Worsley et al. (1992) to control for family-wise error rates. But, this method could be conservative when the number of comparisons is large. Later, a more liberal method, False Discovery Rate (FDR) was proposed by Benjamini and Hochberg (1995) and Genovese et al. (2002). However, the heavy dependence on p-values has been criticized recently in the statistics and science community. Throughout the years, more and more discussion has been surrounding the usage of p-values. Some known statistical drawbacks of using p-values as an evidence metrics include difficulty in interpretation (Hubbard et al. (2003)) and heavy dependence on sample sizes (Greenland et al. (2016); Blume and Peipert (2003); Wasserstein and Lazar (2016)). In addition, the difference between clinical significance and statistical significance (Mark et al. (2016); Ranganathan et al. (2015)) has been gaining attention. With more information presented in the data, it is known that with p-values as an inference tool, even minor differences could result in statistical significance. However, the magnitudes might be too small to be clinically meaningful.

Methods have been suggested to improve upon p-values inference. Yet, most of the improvements have been limited to changing the threshold of significance (Benjamin et al. (2018); Lakens et al. (2018)). Inference tools other than p-values have also been proposed including Likelihood paradigm (Kang et al. (2015); Blume (2002)), Bayes Factors (Bayarri et al. (2016)) and posterior probability (Spiegelhalter et al. (2003)). These methods differ from the common usage of p-values and haven't gathered enough followers to change the clinical practice.

A new method, second-generation p-value (SGPV), was proposed by Blume et al. (2018) to resolve the issues raised in traditional p-values. SGPV preserves some properties of traditional p-values while providing convenient and simple interpretation for non-statisticians. SGPV can be interpreted directly as the proportion of estimates supporting the null hypothesis. Furthermore, SGPV allows the incorporation of clinically meaningful intervals into hypothesis testing. Under the interval testing framework, the Type I error rate is bounded and shrinks to 0 as more information is presented in the data (Blume et al. (2018)). SGPV also promotes good research practice by preventing the post-hoc interpretation of mediocre results. To implement SGPV, an inference interval on a parameter of interest (confidence interval, supportive interval or credible interval) and a clinically meaningful null interval are required. Fortunately, it is convenient to construct such null region in fMRI analysis with the observed data in the cerebrospinal fluid (CSF). Since there is no neuron in CSF, the region should be neurofunctionally null and any detected signals are considered pure noise.

In this study, we propose a novel application of SGPV to overcome the issue often reported in fMRI group inference (Mikl et al. (2008)), i.e., trade-off between more power and inflation of Type I error rates. With SGPV, the Type I error rate can be naturally controlled while gaining enough power as more information is presented. An outline of the paper is as follows. In Section 4.2, we provide an overview of the proposed method and its interpretation. In Section 4.3.1.1, 4.3.2.1 and 4.4.1, we present a simulation study that illustrates the performance of SGPV compared to RFT and FDR. In Section 4.3.1.2, 4.3.2.2 and 4.4.2, we apply the proposed method to real fMRI data analysis. Finally, in Section 4.5, we explore some unanswered questions, summarize the advantages, and discuss some practical issues in the use of SGPV.

4.2 Second-Generation P-values

The novelty of the SGPV lies in its ability to perform interval null hypothesis testing in a straightforward manner. The investigator is able to include in the null space multiple point hypotheses that are scientifically indistinguishable from the traditional point null in that experiment. For example, the null hypothesis for a parameter of interest is usually 0, exactly 0. However, other point hypotheses, i.e., 0.2, 0.1, -0.1, -0.2 may not be scientifically meaningful, especially if the effect can only be estimated to within ± 0.2 units, for example. Of course with a classical p-value, a miniscule effect, such as 0.0001, could even be used to reject the null hypothesis that the effect is zero with large enough sample. Whether or not such an effect is scientifically meaningful is left for post-hoc discussion. What the SGPV does is force investigators to think about this issue before the data are collected (and hence greatly improve the external validity of the experimental findings). If prior clinical knowledge is incorporated when constructing the interval, the inference that follows is both statistically and clinically meaningful.

Let I represent the interval estimate of hypotheses that are supported by the data (e.g. a confidence interval) and H_0 represent the null region. If I = [x,y] where x and y are real numbers and y > x, then the length of I is y - x. The formula for second-generation p-value is defined as:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \ \max\left\{\frac{|I|}{2|H_0|}, 1\right\}$$

where $I \cap H_0$ is the overlap between interval of hypothesis and null region. The first term of the formula denotes the fraction of I that overlaps with H_0 and the second term is a small sample correction factor. From the formula, we can derive that $p_\delta = \frac{|I \cap H_0|}{|I|}$ when $|I| \leq 2|H_0|$ and $p_\delta = \frac{1}{2}\frac{|I \cap H_0|}{|H_0|}$ when $|I| > 2|H_0|$. This shows that when the data are sufficiently precise ($|I| \leq 2|H_0|$), p_δ is just the overlap fraction between I and H_0. When the interval estimate is wide ($|I| > 2|H_0|$), p_δ is reduced to $\frac{1}{2}\frac{|I \cap H_0|}{|H_0|}$, bounded by $\frac{1}{2}$.

Here we provided p_δ under several scenarios.

1. If I is fully contained in H_0, $|I| \leq 2|H_0|$. Then, $p_\delta = \frac{|I \cap H_0|}{|I|} = 1$.

2. If I is fully outside of H_0, $\frac{|I \cap H_0|}{|I|} = 0$. Therefore, regardless of the relationship between $|I|$ and $|H_0|$, $p_\delta = 0$.

3. If I overlaps with H_0 and $|I| \leq 2|H_0|$, p_δ equals to the fraction of I that overlaps with H_0.

4. If I overlaps with H_0 and $|I| > 2|H_0|$, p_δ equals to the fraction of $\frac{1}{2}$ times the overlap fraction.

The interpretation for SGPV is straightforward approximately the proportion fraction of data-supported estimates that are also supporting the null hypothesis. When $p_\delta = 1$, the data-supported estimates only support the null hypothesis. In this case, the data only support the null. When $p_\delta = 0$, the data-supported estimates only support hypotheses that are not null. Here, the data can be said to support only alternative hypotheses, the equivalent of rejecting the null hypothesis in hypothesis testing. When $0 < p_\delta < 1$, the data-supported estimates support both null and alternative hypothesis. Here the data are properly called inconclusive. The magnitude of the SGPV reflects the degree to which the inconclusive data are leading. It is this option that matches most closely with hypothesis tests that fail to reject the null hypothesis. The current data, do not preferentially support either hypothesis over the other.

The SGPV is effectively a three-region summary statistic: support for the null region ($p_\delta = 1$), support for the alternative region ($p_\delta = 0$) and inconclusive region ($0 < p_\delta < 1$). In a Neuroimaging analysis, this translates to the voxel being inactive, active, or undetermined/inconclusive. Note the contrast with traditional p-values, which are a two-region summary statistic: significant or active ($p < 0.05$) and inconclusive or undetermined ($p > 0.05$). The ability to capture support for the null hypothesis is an essential advance for SGPVs and a welcome advance for neuroimager who can now distinguish between active, inactive and undetermined regions of the brain.

Depending on the magnitude of p_δ, results in the inconclusive region could have different meanings. For example, in scenarios where p_δ equals 0.5, the hypothesis intervals are usually very wide, neither supporting nor not supporting the null hypothesis. We call this scenario "strictly inconclusive". On the contrary, some results might be near the extreme ends, i.e., $0 < p_\delta < 0.1$ or $0.9 < p_\delta < 1$. Blume et al. (2019) illustrated an example to report this type of results with a genetic association study. For a result with $0 < p_\delta < 0.1$, the data are suggestive of a meaningful association but are unable to rule out trivial effects. For a result with $0.9 < p_\delta < 1$, the data are suggestive of no association, but not strong enough to rule out meaningful effects. The results near the extreme ends might provide some indication of the association but the evidence is not strong enough to make a definite decision.

In this study, we mimic the decision rules similar to traditional p-values framework for fair comparisons. In addition to the original SGPV (SecondP), we also included the dichotomized version of SGPV (D-SecondP) where voxels with p_δ in the inconclusive region are classified as inactive. We compared the results of D-SecondP with Random Field Theory (RFT) and False Discovery Rate (FDR).

4.3 Methods

4.3.1 Data

4.3.1.1 Simulated data

We validated our proposed inference tool with simulation studies. The data were simulated with spatially and temporally correlated time series at each voxel. After fitting the voxel-specific general linear model, we combined the results and assessed the Type I and II error rates based on different approaches to controlling for multiple comparison: RFT, FDR, SecondP (Second Generation p-values) and D-SecondP (Dichotomous second-generation p-values).

To simulate the spatio-temporal correlation similar to real fMRI data, we adapted similar data generation process as described in Kang et al. (2015) where data were generated following the first order autoregressive model (AR(1)) in a region with spatial dimension 32 × 32 voxels. Data for multiple subjects were simulated in this study. Denote $Y_{\mathrm{v}}(t)$ as the response at a voxel v (v = 1,..., V) and time t (t = 1,..., T) at a single subject level. The model with P stimuli can be expressed as the following for each subject:

$$Y_v(t) = \sum_{p=1}^{P} X_p(t)\beta_v^p + \epsilon_v(t)$$

where X_p denotes the convolution between the p^{th} stimulus impulse function and the Hemodynamic Response Function (HRF). The canonical HRF function from Statistical Parametric Mapping 12 (SPM12) (Friston et al. (2007)) was used to generate $X_p(t)$. We assumed 2 boxcar stimuli for the simulation. The first stimulus was on during [1, $(T/4)$+1] and [$(T/2)$+1, $(3T/4)$+1] and the second is on otherwise. Spatial correlation was implemented on the data via exponential covariance function with the decaying parameter 2 and variance 2.5. The temporal correlation, ϵ_v(t) follows AR(1) process with AR(1) parameter 0.4 and standard deviation 1.5. We assumed that there were two active blocks with 64 voxels and 36 voxels. To compare the methods under various scenarios, we varied the number of time points and number of subjects. To be generalized to real fMRI data, we included T = 64, T = 128 and T = 256 for time points; 5, 10, 20 and 30 subjects and average signal to noise ratios (SNRs) 0.130 (range: 0.0870 ∼ 0.173) (calculated with the formula: effect size / standard deviation). To simulate CSF region consisting of 10 × 10 voxels, the procedure followed the above settings with zero true effect sizes for all parameters related to stimuli in this region.

All simulated data were spatially smoothed with Gaussian kernel filter size at full width, half maximum (FWHM) = 8 mm, except for the scenarios assessing the effects of smoothing kernel size. To explore the effect of smoothing kernel sizes on the methods, we evaluated the performance of RFT, FDR and D-SecondP at sample size = 10 and T= 128 with various degrees of smoothing: FWHM = 0 (Unsmoothed), incorporating information from the immediate neighboring voxels (UseNeighbors), FWHM = 4 and FWHM = 8.

4.3.1.2 Clinical data

For the dataset used in real data application, the data were acquired from a sample of 29 right-handed postmenopausal women between the ages of 45 - 75 with major depressive disorder (MDD). The participants were asked to perform emotion dot probe (EDP) task, a spatial attention task that measures attention bias (Kimonis et al. (2006); Muñoz Centifanti et al. (2013)). Trials of the EDP consisted of a fixation cross presented in the middle of the screen, followed by a brief presentation of a picture pair with one image each on the right and left of the screen. After the picture presentation, a target (asterisk) appeared either on the right or left of the screen (replacing one of the images) and the participant was instructed to indicate by finger button press the side of the screen on which the target appeared as quickly as possible. These images included neutral, positive, and negative (threat and distress) images. The EDP was adapted for fMRI and run as an event related design with three trial types relevant to the presented data: neutral-neutral pair, neutral-negative pair and neutral-positive pair. There were 5 stimuli measured in the study and 6 experimental related covariates. We are interested in the contrast among two of the stimuli, measuring the brain activity difference between pressing on the button and seeing the asterisk when presented negative images.

Participants were scanned on a Philips 3.0 Tesla Achieva scanner, with eight channel head coil. Each subject received a Sagittal T1-weighted 3D Turbo Field Echo Sensitivity Encoding (TR = 9.9 ms; TE = 4.6 ms; a flip angle of 8 degrees) as well as Blood Oxygen Level Dependent (EpiBOLD) functional sequence during the EDP with transverse orientation (TR = 2500 ms; TE = 35 ms; flip angle = 90 degrees). More details about the EDP task and imaging acquisition can be found in Albert et al. (2017). The data were preprocessed using FSL and MATLAB version R2019 (MATLAB (2019)). The preprocessing steps included realignment of the functional runs and correction for bulk-head motion, co-registration of functional and anatomical images for each participant, segmentation of the anatomical image, normalization of the anatomical and functional images to the standard Montreal Neurological Institute (MNI) template. Additionally, the scans were segmented into CSF region. The preprocessed functional images had a voxel size of 2 × 2 × 2 mm and spatially smoothed with Gaussian kernel FWHM = 8 mm.

4.3.2 Statistical Analysis

4.3.2.1 Simulated data

In this simulation, the null hypothesis of interest is H_o: $\beta^2 - \beta^1 = 0$ at each voxel. For FDR procedure, the 2 sided p-values were computed and controlled at 0.05. For RFT, the threshold was set at 3.46, controlling for 2-sided error probability at 0.05. For SecondP and D-SecondP, the clinically null region was deemed at ± (interquartile range (IQR) of $\beta^2 - \beta^1$ of voxels in the CSF region) / 6. The length of this interval was relatively constant throughout all scenarios. To explore the behavior and make fair comparisons between the methods as sample size increases, the length of the interval, i.e., 2 × IQR/6, was found empirically where the corresponding Type I error rates for D-SecondP were similar to FDR at sample size = 5 for all time points.

4.3.2.2 Clinical application

In the analysis, we included the 5 stimuli and 6 covariates measured. Denote the stimuli of interest, D1 and D2. The corresponding regression coefficients are β_v^{D1} and β_v^{D2}. The null hypothesis of interest is H_o: $\beta_v^{D1} - \beta_v^{D2} = 0$. Similar to the simulation settings, the FDR and RFT were controlled for 2-sided error probability at 0.05. The RFT threshold was calculated to be 4.84 (Worsley et al. (1996); Tierney et al. (2016)). For SecondP and D-SecondP, the clinically null region was set at ± (IQR of $\beta_v^{D1} - \beta_v^{D2}$ of voxels in the CSF region) / 6, i.e., [-0.1, 0.1].

To assess the consistency and robustness of each method, we first explored the proportion of voxels categorized as active by each method in the whole brain or relative to the full sample size (N = 29) with various sample sizes. If too few or too many voxels are categorized as active by a method, the method could be either highly conservative or liberal. If there's a large decrease in proportion of active voxels relative to the full sample size (N = 29), then the method can be considered as non-robust. Further, without knowing the ground truth, another way to evaluate the performance of each inference method is a data decimation approach (Zhou et al. (2019); Yang et al. (2014)). According to each inference method, a voxel was deemed "active" or "inactive" based on the activation established with full sample size (N = 29). We treated the result at full sample size as the truth for the data decimation approach. For each smaller sample size, 500 random samples were selected for the analysis, e.g., for N = 10, 500 random combinations were selected out of $\binom{29}{10}$ combinations. The results from smaller sample sizes were compared to the "truth" to estimate the average error rates. Although this method relies on the results from the full sample size and might not give any indication of the ground truth, data decimation allows one to evaluate the ability of a method to reproduce the results at larger sample size and examine the behavior of each method as the sample size decreases. In general, a method that can most closely reproduce the results at full sample size is preferred.

4.4 Results

4.4.1 Simulation study

The simulation results were summarized in Figure to Figure . Figure illustrates the Type I and II error rates of different methods under various time lengths and sample sizes. A key feature of the SGPV method is that it distinguishes between results that are inconclusive and results that support the null, unlike traditional methods which group these two into a single ąğfail to rejectąĺ or ąğnon-significantąĺ region). This leads to difficulty in defining the Type II error rate. By excluding the voxels in the inconclusive region for Type II error rate calculations, the Type II error rates remain the lowest throughout all scenarios for SecondP.

To make fair comparisons with FDR and RFT, D-SecondP was used for comparison where the voxels in the inconclusive regions were deemed as inactive. At T = 64, FDR has the leverage of having smaller Type II error rates throughout most sample size scenarios, followed by the D-SecondP method. The RFT method is the most conservative with the largest Type II error rates. However, as the time length increases, Type II error rates for D-SecondP become similar to FDR while the rates for RFT remain higher. For all methods presented, as time length and sample size increase, Type II error rates decrease dramatically.

At sample size = 5, FDR and D-SecondP share similar Type I error rates. RFT has the highest Type II error rates and the lowest Type I error rates at the same time, reflecting typical Type I and II error rates trade-off. At this sample size, the Type I error rates do not change much with varying time length for all methods. As sample size increases, obvious increases in Type I error rates for FDR and RFT are observed and the error rates rise the most for FDR. Furthermore, as time length increases, the slopes between Type I error rates and sample size become steeper for both FDR and RFT but remain relatively constant for D-SecondP. It is worth noting that RFT starts off with the lowest Type I error rate at sample size = 5 but reaches to the same

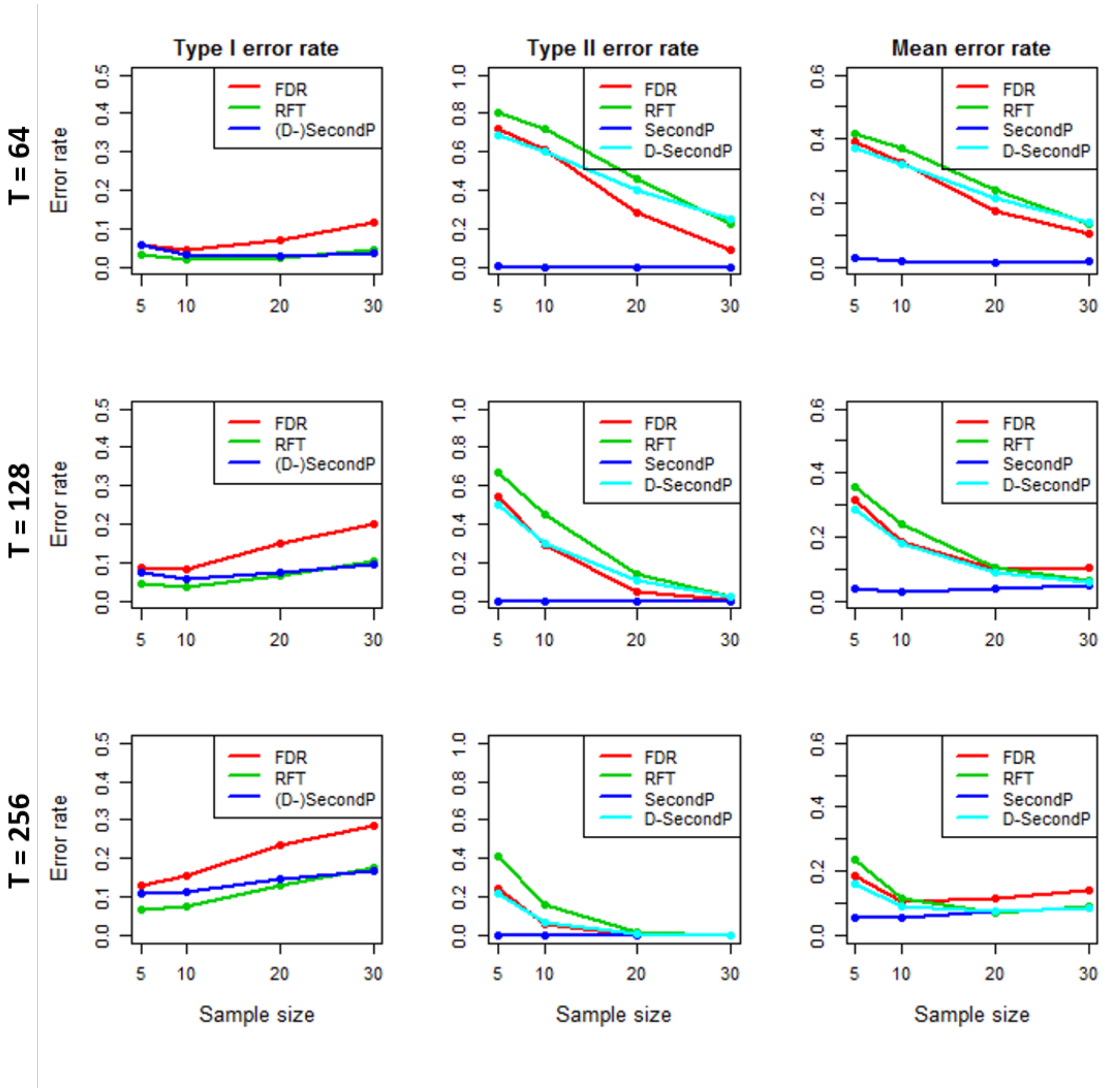


Figure 4.1: Results of the simulation study based on 300 repetitions with different sample sizes at 5, 10, 20 and 30. Rows represent the different time series at T = 64, 128 and 256. Columns represent the different error rates. The first column corresponds to the Type I error rates among the truly null regions (the average number of false positives divided by the total number of voxels in the null regions). The second column corresponds to the Type II error rates among the truly active regions (the average number of false negatives divided by the total number of voxels in the active regions). The third column is the average of the first two columns. Different inference methods are indicated with different colors: FDR (red), RFT (green), SecondP (blue) and D-SecondP (turquoise). Note that SecondP and D-SecondP shares the same Type I error rate.

level as D-SecondP at all time lengths when sample size reaches 30.

One way to evaluate the joint performance of Type I and II error rates is to compute the mean error rates, the average between Type I and II error rates. Due to the conservative nature of RFT, the mean error rates for RFT are the highest in most scenarios. At T = 64, large decrease in Type II error rates for FDR outweighs the increase in Type I error rates, resulting in the lowest mean error rates. However, as time length increases, the dramatic increase in FDR Type I error rates leads to the highest mean error rates with increasing sample sizes. The mean error rates for D-SecondP become the lowest with steady Type I error rates compared to RFT and FDR as time length and sample size increase.

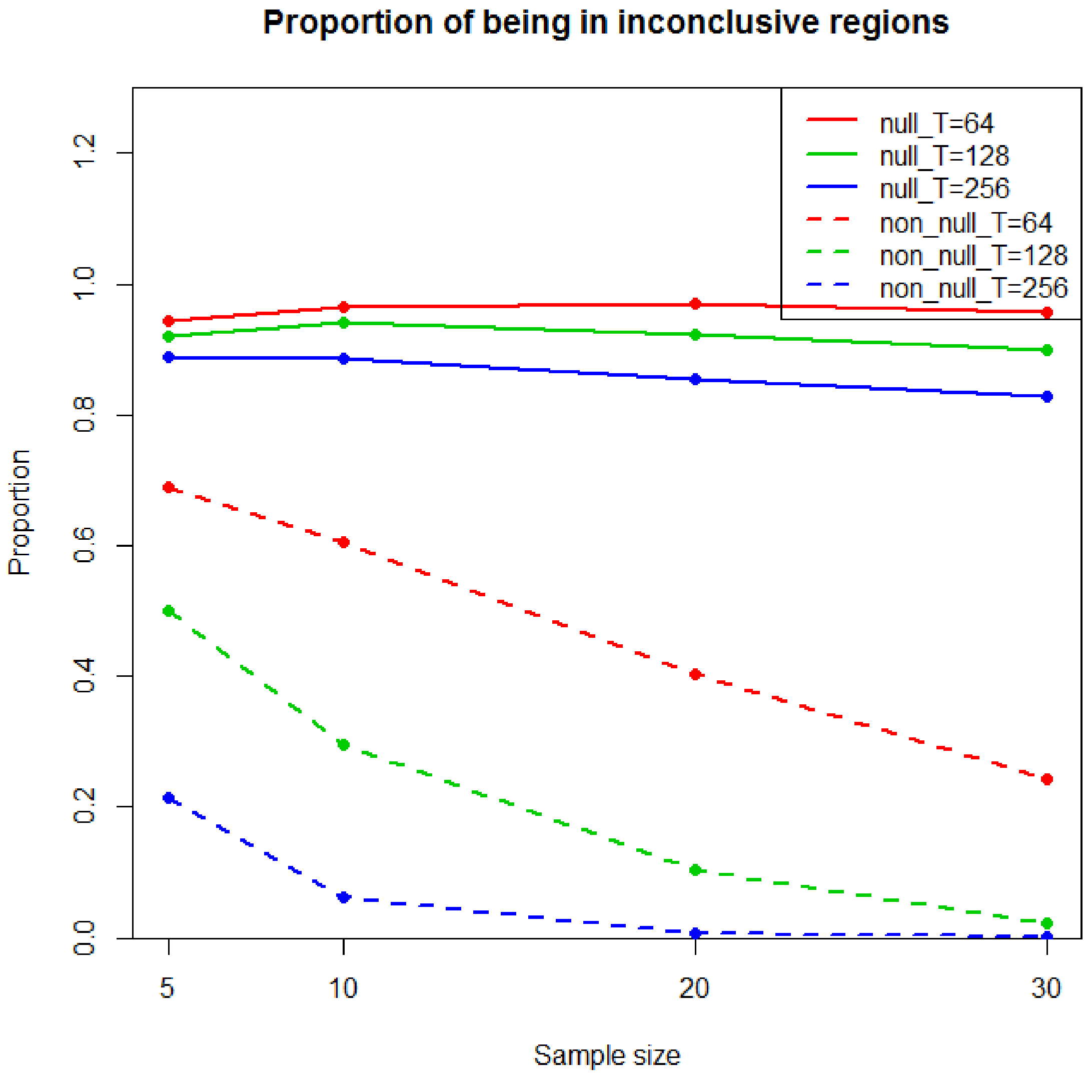


Figure 4.2: Results of the simulation study based on 300 repetitions in terms of proportion of voxels in the inconclusive region where the SGPV for the voxel is greater than 0 and less than 1. The average proportion was calculated as the average number of voxels in the inconclusive region divided by the total number of voxels in the null or the non-null region. The solid lines correspond to the null regions and dashed lines correspond to the non-null regions with T = 64 (red), T = 128 (green) and T = 256 (blue).

To assess the behavior of voxels in the inconclusive ($0 < p_\delta < 1$) and strictly

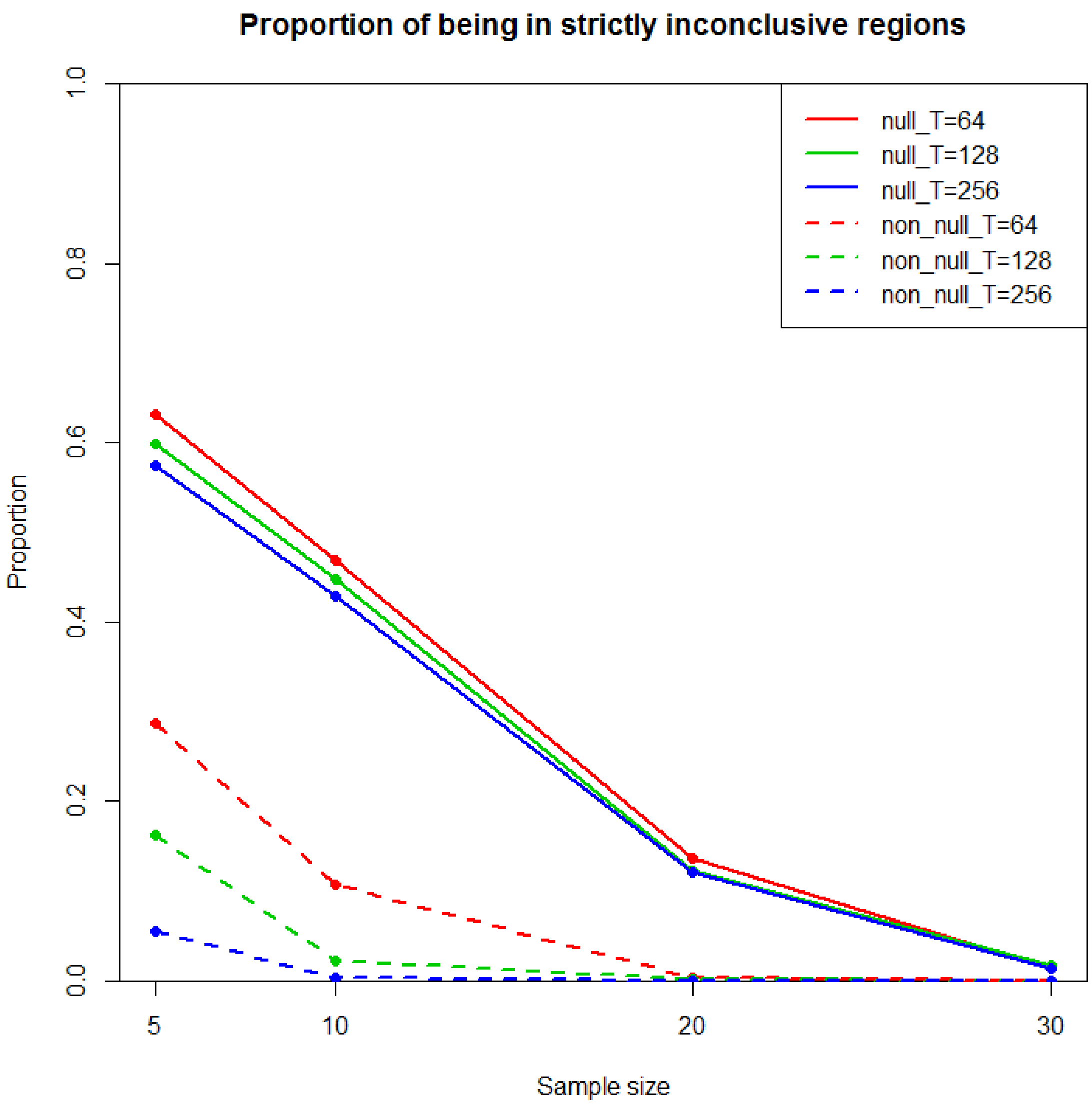


Figure 4.3: Results of the simulation study based on 300 repetitions in terms of proportion of voxels in the strictly inconclusive region where the SGPV for the voxel is 0.5. The average proportion was calculated as the average number of voxels in the strictly inconclusive region divided by the total number of voxels in the null or the non-null region. The solid lines correspond to the null regions and dashed lines correspond to the non-null regions with T = 64 (red), T = 128 (green) and T = 256 (blue).

inconclusive ($p_\delta = 0.5$) regions, the average proportions of voxels in these two regions at various time lengths and sample sizes are summarized in Figure 4.2 and Figure 4.3 (note that strictly inconclusive region is a subset of inconclusive region). If a voxel belongs to one of the two active blocks described in Section 4.3.1.1, it is in the non-null region. If a voxel doesn't belong to one of the two active blocks described in

Section 4.3.1.1, it is in the null region. In the non-null regions (dashed lines), as time length and sample size increase, the average proportions of voxels in both regions decrease as shown in Figure 4.2 and Figure 4.3. In addition, with more information presented, the average proportion converges to 0 faster (at T = 128, the average proportion drops to 0 at sample size = 30 while at T = 256, the average proportion drops to 0 at sample size = 20 (Figure 4.2)). The average proportion of voxels in the strictly inconclusive null region decreases as sample size increases (Figure 4.3) but remains relatively constant in the inconclusive null region (Figure 4.2). The effect of time length on the average proportions is less noticeable in null regions.

Lastly, to explore the spatial smoothing effect in all methods, we evaluated the error rates under different smoothing methods at T = 128 and sample size = 10 (Figure 4.4). All methods perform the worst in terms of mean error rates when there's no smoothing implemented. The performance becomes better after taking into account underlying spatial correlation, resulting in lower Type II error rates. RFT obtains the lowest Type I and the highest Type II error rates with all smoothing methods. Both the Type I and II error rates are the lowest for D-SecondP under all smoothing methods. For RFT, FDR and D-SecondP, the Type I error rates increase as the smoothing kernel size increases and the error rate increases the most for FDR.

4.4.2 Clinical application

The average proportions of active voxels in the whole brain and relative to full sample with respect to different sample size are summarized in the top row plots of Figure 4.5. Although RFT has the lowest proportion of active voxels in the whole brain throughout all sample sizes (around 3%), the average proportion of active voxels drops to 40% relative to the full sample size (N = 29) at sample size = 10. The average proportions of active voxels change the most for FDR. The average proportion drops from 40% to 10% in the whole brain and only 30% of the active voxels remain relative to the full data when sample size drops to 10. In contrast, D-SecondP maintains the average proportions more steadily. The proportion drops from 20% to 10% in the whole brain and 60% of the active voxels remain relative to the full data when sample size drops to 10, indicating that the inference based on D-SecondP is more robust to decreasing sample size compared to the other methods.

The observation from the top row plots in Figure 4.5 can be partially visualized in Figure 4.6. The RFT has the least activation among all methods compared at both sample sizes. This conservative nature was also observed in Figure 4.5 and simulation

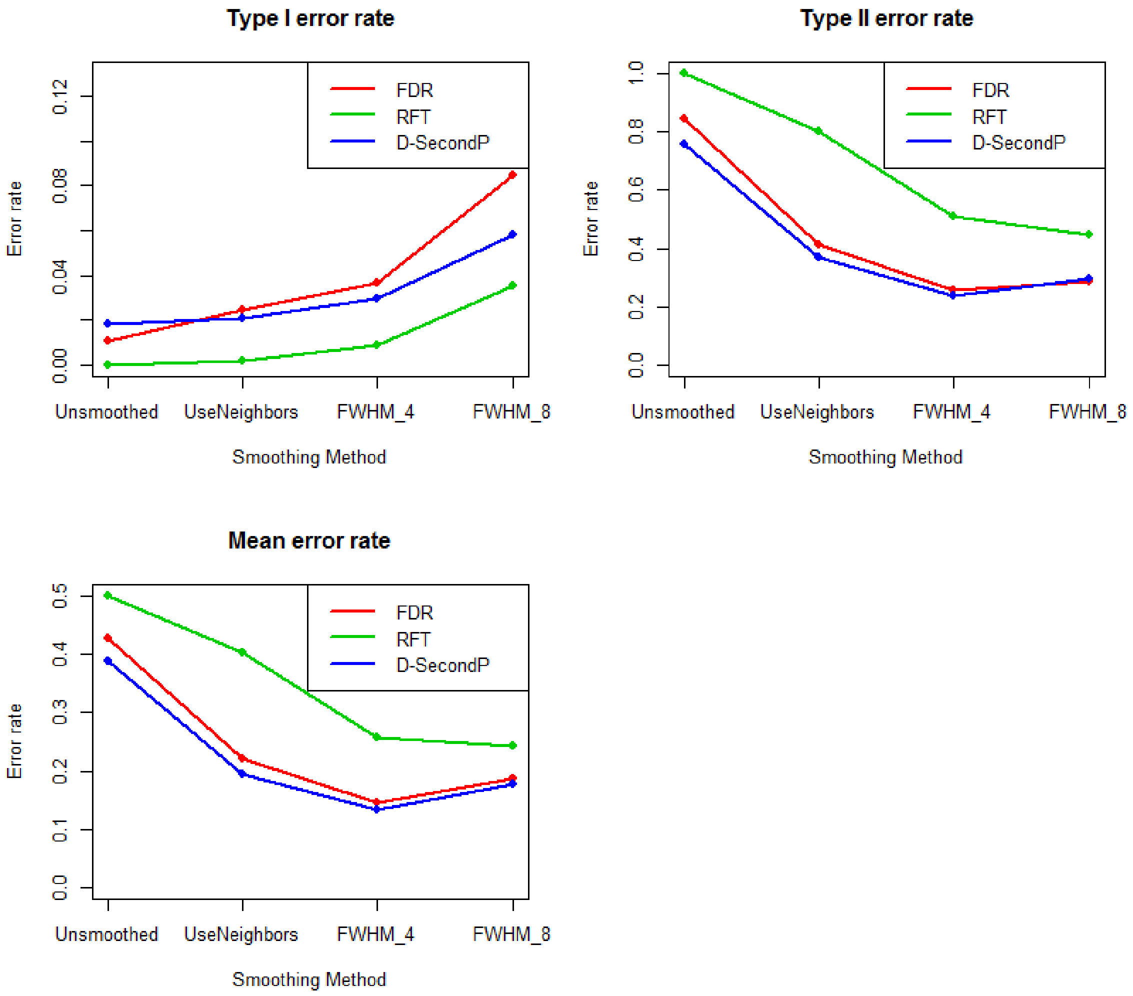


Figure 4.4: Simulation results for FDR (red), RFT (green), D-SecondP (blue) at T = 128 and sample size = 10. Type I, II and mean error rates are shown with varying smoothing methods: unsmoothed, UseNeighbors, FWHM=4 mm and FWHM = 8 mm.

results. At sample size = 15, the activation patterns are very similar between FDR and D-SecondP. However, at full sample size, although more active voxels are observed for both FDR and D-SecondP compared to sample size = 15, the increase for FDR is more drastic than D-SecondP. In our study, the voxels are deemed active in most of the latter half of the brain with FDR, covering larger areas in and between ROIs. The boarders between different ROIs become unclear.

The results of data decimation are shown in the bottom row plots of Figure 4.5. Treating the activation patterns at full sample size the truth, RFT has the lowest Type I error rates compared to FDR and D-SecondP throughout all sample sizes. As expected, it also has the highest Type II error rates. At sample size = 10, the

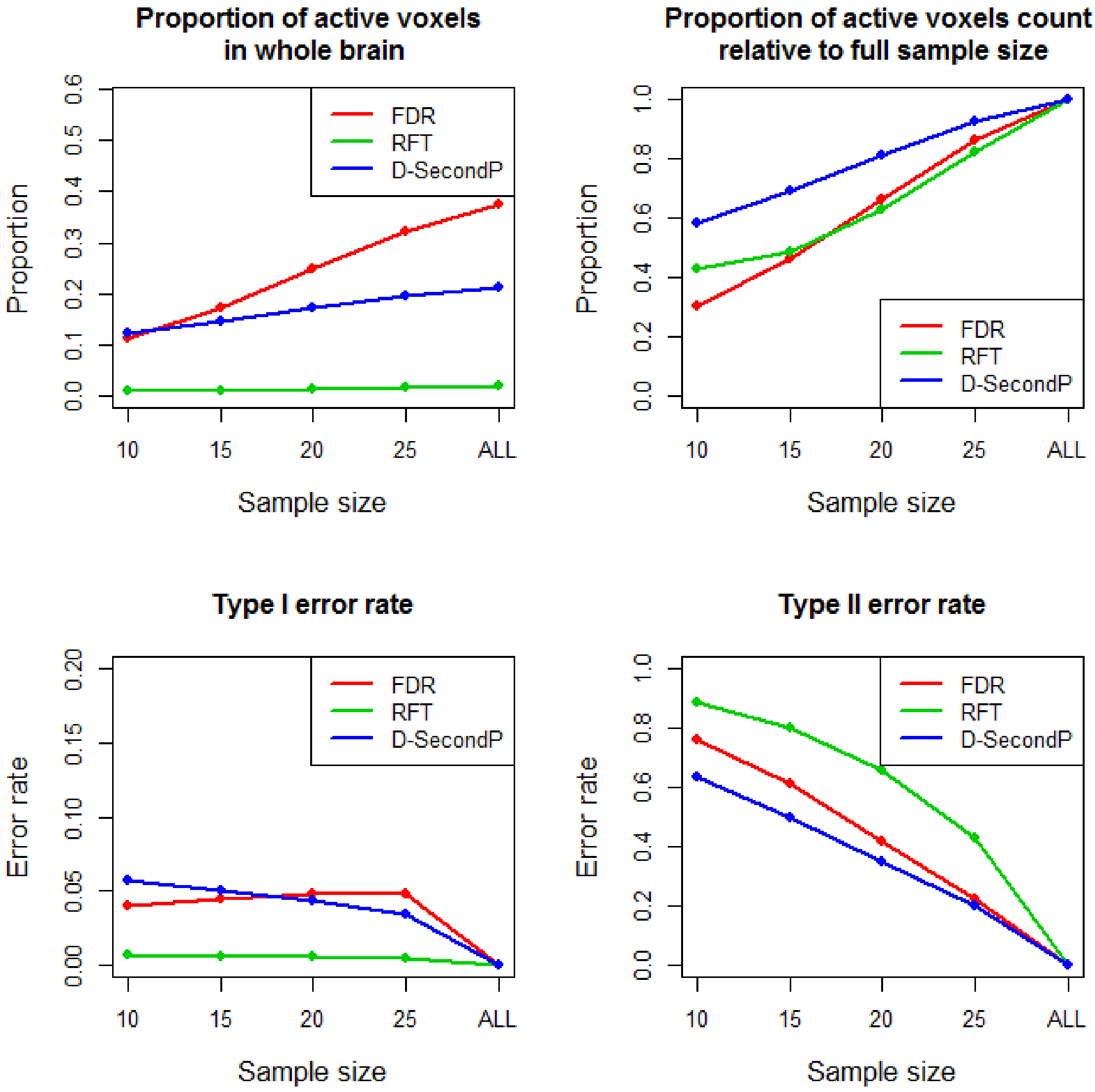


Figure 4.5: Real data analysis comparing FDR (red), RFT (green) and D-SecondP (blue). The top left image corresponds to the average proportion of active voxels in the whole brain with varying sample sizes. The average proportion was calculated as the average number of active voxels divided by the total number of voxels used in the analysis. The top right image corresponds to the average proportion of active voxels relative to the full sample size. The average proportion was calculated as the average number of active voxels at each sample size divided by the number of active voxels at full sample size. Results for Data decimation are shown in the bottom row. The voxels patterns at full sample size were treated as the truth. Type I error rate was calculated as the average number of false positives at each sample size divided by the total number of voxels in the null regions at full sample size. The Type II error rate was calculated as the number of false negatives at each sample size divided by the total number of voxels in the non-null regions at the full sample size.

error rate is as high as 90%. D-SecondP has slightly lower Type I error rate at larger sample sizes compared to FDR. In addition, throughout all sample sizes, D-SecondP

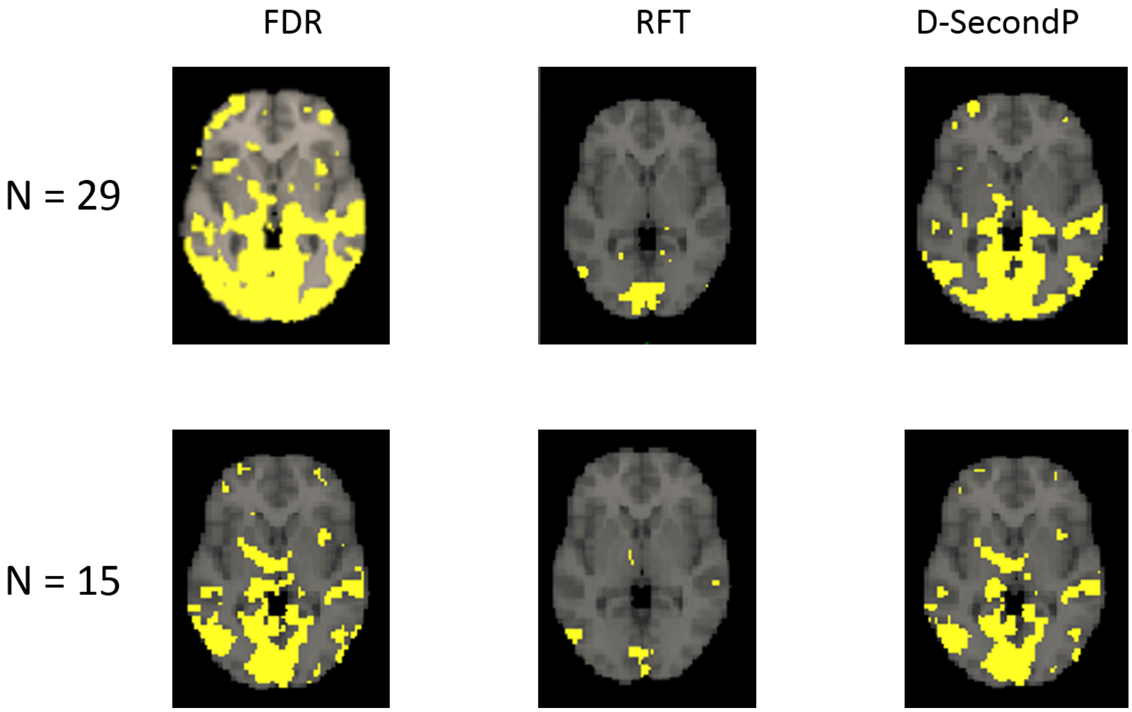


Figure 4.6: Data decimation results of activation maps for the 37th axial slice. The yellow blobs indicate activated areas resulted from FDR (first column), RFT (second column) and D-SecondP (third column). The first row corresponds to the activation maps with full sample size at 29. The second row corresponds to the activation maps with a randomly selected sample of size 15.

has lower Type II error rates than FDR. With lower Type II error rate and decent Type I error rate control, D-SecondP is the most robust method against decreasing sample size among the methods compared in this study.

4.5 Discussion

Current fMRI inference still heavily relies on p-values. Drawbacks of p-values have been discussed in both statistical and scientific fields. The newly proposed method, SGPV, offers convenient and simple interpretation to non-statisticians. The goal of this paper is to evaluate SGPV as an inference tool in fMRI analysis compared to other commonly used methods under various experimental settings. We assessed the behaviors of SGPV, RFT and FDR under different time lengths and sample sizes with simulated and real fMRI data. In both simulation and data decimation using real fMRI data, SGPV shows better performance in terms of average test error rates than both RFT and FDR. With the interval null, SGPV is not only able to decrease the Type II error rate as more data are presented compared to RFT but also attenuate the inflation of Type I error rate compared to FDR (Figure 4.1). Further, we have shown that SGPV is more robust to decrease in sample size than the conventional approaches via a data decimation study (Figure 4.5). However, two questions remained unanswered by the results presented. First, what was the reason for the discrepancy observed in Figure 4.2 and Figure 4.3 where the proportion of voxels in strictly inconclusive region ($p_\delta = 0.5$) shrinks to 0 (Figure 4.3) but stays relatively constant in the null inconclusive region ($0 < p_\delta < 1$) as sample size increases (Figure 4.2)? Second, what was the reason for better control of Type I error rate with D-SecondP over other methods? To further address these questions, we specifically looked into the behaviors of the voxels in the null region with simulated data presented in Section 4.4.1 at T = 128 and sample size = 5 and 30. We denote null voxels that are close to the active voxels in the "Neighboring" zone, consisting of 7% of the null region and the remaining 93% of the null voxels in the "Remaining" zone.

From Figure 4.3, as sample size increases, the proportion of voxels in the null strictly inconclusive region decreases to around 0. The hypothesis intervals are usually very wide in the strictly inconclusive region and neither support nor not support the null hypothesis. Once more information is presented, voxels move out of this region and towards support for alternative region ($p_\delta = 0$) or support for the null region ($p_\delta = 1$). We further explore the behavior of these voxels after leaving the strictly inconclusive region. At sample size = 5, the median SGPV in the "Neighboring" zone is 0.32. 24.7% of voxels are in the support for the alternative region ($p_\delta = 0$) and the remaining 75.3% of voxels are in the inconclusive region (33.5% in the strictly inconclusive region; 36.8% with $0 < p_\delta < 0.5$ and 5% with $0.5 < p_\delta < 1$). When sample size increases to 30, the median SGPV drops to 0.009. 71.5% of voxels are in the support for the alternative region ($p_\delta = 0$) and the remaining 28.5% of voxels are in the inconclusive region (0% in the strictly inconclusive region; 27.1% with $0 < p_\delta$ < 0.5 and 4% with $0.5 < p_\delta < 1$). There is a large increase in the proportion of voxels in the support for the alternative region ($p_\delta = 0$) but only a minor difference in the proportion of voxels with $0 < p_\delta < 0.5$. We can infer that most voxels, initially in the strictly inconclusive region, move towards the support for the alternative region (p_δ $= 0$) and large portion of the voxels with $0 < p_\delta < 0.5$ moves into the support for the alternative region ($p_\delta = 0$). Conceptually, voxels in the null region should have null effect. As sample size increases, null voxels should move towards the support for the null region ($p_\delta = 1$). One simple explanation is that these voxels become falsely active due to spatial smoothing (Mikl et al. (2008)). During spatial smoothing, non-zero false effects are artificially added to the null voxels by weighing in neighboring active voxels. Voxels in the "Neighboring" zone are highly close to the active voxels and became almost indistinguishable from the truth. At sample size = 5, around 26% of voxels in this zone are mistakenly categorized as active by FDR and D-SecondP. Due to the conservative nature of RFT, these voxels are still correctly classified as inactive by RFT. The result corresponds to the lowest Type I error rates observed for RFT at sample size = 5 and similar Type I error rate for FDR and D-SecondP (Figure 4.1). Although voxels in this zone only consists of 7% of the null region, as sample size increases to 30, more than 70% of voxels are mistakenly categorized as active by FDR, D-SecondP and RFT.

Comparing to voxels in the "Neighboring" zone, voxels in the "Remaining" zone are located further away from active voxels. At sample size = 5, the median SGPV in the "Remaining" zone is 0.5. 94% of voxels are in the inconclusive region (61.8% in the strictly inconclusive region; 22.1% with $0 < p_\delta < 0.5$ and 10% with $0.5 < p_\delta < 1$). When sample size increases to 30, the median SGPV in the "Remaining" zone is 0.58. 94.4% of voxels are in the inconclusive region (1.7% in the strictly inconclusive region; 30.4% with $0 < p_\delta < 0.5$ and 62.3% with $0.5 < p_\delta < 1$). The proportion of voxels in the inconclusive region does not change. This shows that after leaving the strictly inconclusive region, voxels remain in the inconclusive region where data are not strong enough to make a definite decision for these voxels as sample size increases. Both proportions of voxels with $0 < p_\delta < 0.5$ and $0.5 < p_\delta < 1$ increase. This implies that heterogeneity exists among these migrated voxels. Some voxels move towards the support for the null region ($p_\delta = 1$) as expected and end up with $0.5 < p_\delta < 1$. These voxels are located the furthest away from active voxels with negligible influence from spatial smoothing. Others move towards the support for the alternative region ($p_\delta = 0$) and have $0 < p_\delta < 0.5$. These voxels are located closer to the "Neighboring" zone compared to voxels with $0.5 < p_\delta < 1$. Although voxels with $0 < p_\delta < 0.5$ are less affected by spatial smoothing compared to voxels in the "Neighboring" zone, most of the added false effects are still influential. We believe that with wider range of null values in the null interval (H_0), most of the false effects in the "Remaining" zone are captured as part of the interval null making these voxels distinguishable from the truth and are kept in the inconclusive region. In contrast, the false effects are more likely to be detected by methods under point null hypothesis testing framework. As sample size increases, the influence from false effects becomes worse. More false active voxels are then identified by FDR. The Type I error rates for FDR start from similar values as D-SecondP at sample size = 5 to higher values at sample size = 30, leading to a steeper increase in Type I error rates compared to D-SecondP as shown in Figure 4.1. In our simulation study, RFT is slightly conservative. Voxels in the "Remaining" zone are still correctly classified as inactive by RFT. Therefore, at sample size = 30, the Type I error rates for RFT and D-SecondP are similar (Figure 4.1). The Type I error rates for RFT start from almost 0 at sample size = 5 to similar values as D-SecondP at sample size = 30, leading to a steeper increase of Type I error rates compared to D-SecondP (Figure 4.1). However, in real fMRI data analysis with larger number of comparisons, RFT is highly conservative with low power and low Type I error rate (Figure 4.5 and Figure 4.6). To sum up, we have explained the discrepancy in Figure 4.2 and Figure 4.3 and further illustrate that by incorporating the inconclusive region and interval testing, SGPV is able to discern the true active voxels with better control of false positives over other common methods.

Another desired property of SGPV is the ability to incorporate clinical knowledge into hypothesis testing. This allows the investigator to draw both statistically and clinically meaningful inference. In addition, this property also helps promote good statistical practice by preventing post-hoc interpretation. Nevertheless, constructing the clinical region can be subjective. The interval chosen in this study should serve as a guide. Prior knowledge in clinical practice and data noise patterns are required when choosing an appropriate interval range and demand further research. With a fixed and pre-determined clinically null region, we have shown that the SGPV clearly outperforms the conventional approaches in various scenarios while offering unique strengths that differ from the traditional hypothesis testing. To make the usage of SGPV more generalizable in fMRI analysis, an R shiny app is currently being developed for easy visualization. The results can be exported in Neuroimaging Informatics Technology Initiative (NIfTI) format as inputs to other imaging software like FSL or SPM12.

CHAPTER 5

CONCLUSION

5.1 Summary

This dissertation aims to bring out some complexities one might face when handling large dimensional data while novel methods can be computationally burdensome. We focus on providing alternative, ready-to-use methods that target to solve specific issues in the data more efficiently. In Chapter 2, we introduced PheWAS in EMR data and proposed to estimate relative risk instead of odds ratios to overcome the need for exclusion criteria. In Chapter 3, we thoroughly demonstrated the potential influence spatial smoothing and other experimental factors can have on statistical inference. We also explored an alternative spatial smoothing method. The factors contributing to statistical inference issues raised in Chapter 3 have largely been ignored. Most of the current methods have drawbacks and are not ready to be implemented in voxel-wise analysis. In order to accommodate the deficiency with p-values inference framework, we introduced a novel inference framework that separates from the traditional point hypothesis testing for identification of active brain regions related to specific task of interest. This technique is shown to outperform traditional frequentist methods in terms of balance between Type I and II error rates.

5.2 PheWAS Analysis Improvement in EMR Data

Section I (Chapter 2) of this dissertation focused on better understanding of PheWAS analysis and provided an alternative method that overcomes the limitation raised in this chapter. We introduced the purpose of PheWAS and limitation encountered when performing logistic regression analysis. The main difficulty in PheWAS is to correctly classify the case statuses of the phenotypes in EMR data. While the standard method tried to achieve the goal, the accuracy of the manually-compiled exclusion criteria lists for classifying controls population cannot be guaranteed without extensive data curation process. The inefficiency hinders PheWAS from being extended to handle larger-scale phenome construction agnostic analysis of phenotypes that preserve more disease-related clinical information. Other methods tried to improve upon the classification issue by direct estimation of the misclassification rates. However, in addition to being computationally burdensome, the prevalence of the phenotypes in EMR varies, often leading to model convergence failure. We demonstrated via simulation and real data application that without accurately classifying the controls population, large biases can occur when analyzing prevalent diseases. The desired nature of relative risks overcomes the need for exclusion criteria lists and provides efficient, reliable, and unbiased estimation. Currently, there are 13000 ICD 9 codes and 68000 ICD 10 codes. We can expect more and more clinical information to be provided in the future versions of the billing codes. By allowing PheWAS to bypass the exclusion criteria lists, we are able to efficiently and reliably extend PheWAS to adapt to larger-scale phenome construction agnostic analyses of phenotypes which contain more disease-related clinical information.

5.3 Task-induced fMRI Data Analysis via Second-generation P-values

Section II (Chapter 3-4) of this dissertation focused on task-induced fMRI data analysis. fMRI data are noisy and require lots of preprocessing steps before analysis. However, less attention was paid to investigate the influence of the steps on the statistical inference. In Chapter 3, we explored the spatial smoothing and demonstrate the influence along with experimental factors on the inference. We have shown that with traditional frequentist techniques, large trade-off of Type I error rates for power is observed when more information is added in the analysis. More inactive voxels were misclassified as active voxels due to the smoothing procedure. However, other approaches are mostly computationally burdensome and solely focus on controlling the Type I error rates, ignoring the balance between Type I and II error rates. In order to address the critical issue, we introduced second-generation p-values by bringing in the interval null hypothesis testing in Chapter 4. By allowing the interval null, the falsely active voxels can be contained in the null and allow the true active voxels to be discern. We have shown in simulation and data analysis that the proposed technique allows for steady control of Type I error rates while obtaining enough power, resulting in improved inference compared to traditional methods.

5.4 Summary Contributions

- We proposed to estimate relative risk in PheWAS analysis instead of odds ratio. It overcomes the traditional difficulties in classification of the controls population. The efficient and ready-to-use method allows PheWAS to be extended to larger-scale phenomes analyses.

- We thoroughly explored and demonstrated the influence of degree of spatial smoothing and experimental factors on the statistical inference. We also extended the smoothing method to maximum likelihood estimates.

- We provided methodology for the detection of active voxels related to the task of interest using SGPV applied to task-induced fMRI data. The use of our methods allows for better balance between Type I and II error rates in identifying activated regions as demonstrated on simulated and clinical data.

REFERENCES

Albert, K., Gau, V., Taylor, W. D. and Newhouse, P. A. (2017), Attention bias in older women with remitted depression is associated with enhanced amygdala activity and functional connectivity, *Journal of Affective Disorders* **210**, 49–56.

Bartés-Serrallonga, M., Serra-Grabulosa, J. M., Adan, A., Falcón, C., Bargalló, N. and Solé-Casals, J. (2014), Smoothing FMRI data using an adaptive Wiener filter, *in* 'Studies in Computational Intelligence', Vol. 577, Springer Verlag, 321–332.

Bayarri, M. J., Benjamin, D. J., Berger, J. O. and Sellke, T. M. (2016), Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses, *Journal of Mathematical Psychology* **72**, 90–103.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. and Johnson, V. E. (2018), *Redefine statistical significance*, Vol. 2, Nature Publishing Group.

Benjamini, Y. and Hochberg, Y. (1995), Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple, Technical Report 1.

Bennett, C. M. and Miller, M. B. (2010), How reliable are the results from functional magnetic resonance imaging?, *Annals of the New York Academy of Sciences* **1191**(1), 133–155.

Blume, J. D. (2002), Likelihood methods for measuring statistical evidence, *Statistics in Medicine* **21**(17), 2563–2599.

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D. and Greevy, R. A. (2018), Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses, *PLOS ONE* **13**(3), e0188299.

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R. and Dupont, W. D. (2019), An introduction to second-generation p-values, *The American Statistician* **73**(sup1), 157–167.

Blume, J. and Peipert, J. F. (2003), What Your Statistician Never Told You about P-Values, *Journal of the American Association of Gynecologic Laparoscopists* **10**(4), 439–444.

Bullmore, E., Rabe-Hesketh, S., Morris, R., Williams, S., Gregory, L., Gray, J. and Brammer, M. (1996), Functional Magnetic Resonance Image Analysis of a Large-Scale Neurocognitive Network, *NeuroImage* **4**(1), 16–33.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F. and Parkinson, H. (2019), The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, *Nucleic Acids Research* **47**(D1), D1005–D1012.

Crofford, L. J. (2013), Use of NSAIDs in treating patients with arthritis., *Arthritis research & therapy* **15 Suppl 3**(Suppl 3), S2.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., Bielinski, S. J., Pendergrass, S. A., Xu, H., Hindorff, L. A., Li, R., Manolio, T. A., Chute, C. G., Chisholm, R. L., Larson, E. B., Jarvik, G. P., Brilliant, M. H., McCarty, C. A., Kullo, I. J., Haines, J. L., Crawford, D. C., Masys, D. R. and Roden, D. M. (2013), Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data., *Nature biotechnology* **31**(12), 1102–10.

Denny, J. C., Bastarache, L. and Roden, D. M. (2016), Phenome-Wide Association Studies as a Tool to Advance Precision Medicine., *Annual review of genomics and human genetics* **17**, 353–73.

Edwards, J. K., Cole, S. R., Troester, M. A. and Richardson, D. B. (2013), Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data., *American journal of epidemiology* **177**(9), 904–12.

Eklund, A., Nichols, T. E. and Knutsson, H. (2016), Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates, *Proceedings of the National Academy of Sciences of the United States of America* **113**(28), 7900–7905.

Firth, D. (1993), Bias Reduction of Maximum Likelihood Estimates, *Biometrika* **80**(1), 27.

Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T. and Penny, W. D. (2007), *Statistical parametric mapping : the analysis of funtional brain images*, Elsevier/Academic Press.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002), Thresholding of statistical maps in functional neuroimaging using the false discovery rate, *Neuroimage* **15**(4), 870–878.

Grant, R. L. (2014), Converting an odds ratio to a range of plausible relative risks for better communication of research findings., *BMJ (Clinical research ed.)* **348**, f7450.

Greenland, S., Stephen, ., Senn, J., Kenneth, ., Rothman, J., Carlin, J. B., Poole, C., Goodman, S. N., Douglas, ., Altman, G., Senn, S. J. and Altman, D. G. (2016), Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations, *Eur J Epidemiol* **31**, 337–350.

Hebbring, S. J., Rastegar-Mojarad, M., Ye, Z., Mayer, J., Jacobson, C. and Lin, S. (2015), Application of clinical text data for phenome-wide association studies (PheWASs)., *Bioinformatics (Oxford, England)* **31**(12), 1981–7.

Hopfinger, J. B., Buonocore, M. H. and Mangun, G. R. (2000), The neural mechanisms of top-down attentional control, *Nature Neuroscience* **3**(3), 284–291.

Hubbard, R., Bayarri, M. J., Berk, K. N. and Carlton, M. A. (2003), Confusion over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing General Confusion Over Measures of Evidence (p's) Versus Errors (a's) in Classical Statistical Testing, *Source: The American Statistician* **57**(3), 171–182.

Kang, H., Blume, J., Ombao, H. and Badre, D. (2015), Simultaneous control of error rates in fmri data analysis, *Neuroimage* **123**, 102–113.

Kang, H., Ombao, H., Linkletter, C., Long, N. and Badre, D. (2012), Spatio-Spectral Mixed-Effects Model for Functional Magnetic Resonance Imaging Data, *Journal of the American Statistical Association* **107**(498), 568–577.

Katanoda, K., Matsuda, Y. and Sugishita, M. (2002), A Spatio-temporal Regression Model for the Analysis of Functional MRI Data, *NeuroImage* **17**(3), 1415–1428.

Kimonis, E. R., Frick, P. J., Fazekas, H. and Loney, B. R. (2006), Psychopathy, aggression, and the processing of emotional stimuli in non-referred girls and boys, *Behavioral Sciences & the Law* **24**(1), 21–37.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S. and Turner, R. (1992), Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation., *Proceedings of the National Academy of Sciences of the United States of America* **89**(12), 5675–9.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S. C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., Debruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., Van Harmelen, A. L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q., Nilsonne, G., De Oliveira, C. L., De Xivry, J. J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I. and Zwaan, R. A. (2018), *Justify your alpha*, Vol. 2, Nature Publishing Group.

Lang, P. J., Bradley, M. M. and Cuthbert, B. N. (1999), International affective picture system (iaps): Technical manual and affective ratings, Technical report, University of Florida.

Lindquist, M. A. and Mejia, A. (2015), Zen and the Art of Multiple Comparisons, *Psychosomatic medicine* **77,2**.

Liu, H. and Zhang, Z. (2017), Logistic regression with misclassification in binary outcome variables: a method and software, *Behaviormetrika* **44**(2), 447–476.

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y. and Sobel, J. D. (2011), Validation data-based adjustments for outcome misclassification in logistic regression: an illustration., *Epidemiology (Cambridge, Mass.)* **22**(4), 589–97.

Mark, D. B., Lee, K. L. and Harrell, F. E. (2016), Understanding the role of P values and hypothesis tests in clinical research, *JAMA Cardiology* **1**(9), 1048–1054.

Marschner, I. C. and Gillett, A. C. (2012), Relative risk regression: reliable and flexible methods for log-binomial models, *Biostatistics* **13**(1), 179–192.

MATLAB (2019), *version 9.6 (R2019a)*, The MathWorks Inc., Natick, Massachusetts.

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M. and Krupa, P. (2008), Effects of spatial smoothing on fMRI group inferences, *Magnetic Resonance Imaging* **26**(4), 490–503.

Muñoz Centifanti, L. C., Kimonis, E. R., Frick, P. J. and Aucoin, K. J. (2013), Emotional reactivity and the association between psychopathy-linked narcissism and aggression in detained adolescent boys, *Development and Psychopathology* **25**(2), 473–485.

Neuraz, A., Chouchana, L., Malamut, G., Le Beller, C., Roche, D., Beaune, P., Degoulet, P., Burgun, A., Loriot, M.-A. and Avillach, P. (2013), Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics, *PLoS Computational Biology* **9**(12), e1003405.

Newlander, S. M., Chu, A., Sinha, U. S., Lu, P. H. and Bartzokis, G. (2014), Methodological improvements in voxel-based analysis of diffusion tensor images: Applications to study the impact of apolipoprotein e on white matter integrity, *Journal of Magnetic Resonance Imaging* **39**(2), 387–397.

Ogawa, S., Lee, T. M., Kay, A. R. and Tank, D. W. (1990), Brain magnetic resonance imaging with contrast dependent on blood oxygenation., *Proceedings of the National Academy of Sciences of the United States of America* **87**(24), 9868–72.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H. and Ugurbil, K. (1992), Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging., *Proceedings of the National Academy of Sciences* **89**(13), 5951–5955.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. and Tanaka, T. (2002), Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction, *Nature Genetics* **32**(4), 650–654.

Pajula, J. and Tohka, J. (2014), Effects of spatial smoothing on inter-subject correlation based analysis of FMRI, *Magnetic Resonance Imaging* **32**(9), 1114–1124.

Ranganathan, P., Pramesh, C. and Buyse, M. (2015), Common pitfalls in statistical analysis: Clinical versus statistical significance, *Perspectives in Clinical Research* **6**(3), 169.

Slotnick, S. D. (2017), Cluster success: fmri inferences for spatial extent have acceptable false-positive rates, *Cognitive Neuroscience* **8**(3), 150–155. PMID: 28403749.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M. and Matthews, P. M. (2004), Advances in functional and structural MR image analysis and implementation as FSL, *NeuroImage* **23**, S208–S219.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2003), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley & Sons, Ltd, Chichester, UK.

Stock, J. and Watson, M. W. (2008), Heteroskedasticity-Robust Standard Errors for Fixed Effects Regression, *Econometrica* **76**.

Strappini, F., Gilboa, E., Pitzalis, S., Kay, K., Mcavoy, M., Nehorai, A. and Snyder, A. Z. (2016), Adaptive Smoothing Based on Gaussian Processes Regression Increases the Sensitivity and Specificity of fMRI Data.

Tierney, T. M., Clark, C. A. and Carmichael, D. W. (2016), Is bonferroni correction more sensitive than random field theory for most fmri studies?, *arXiv* .

Wang, D. J., Zhu, D. H., Fan, D. J., Giovanello, D. K. and Lin, D. W. (2013), Multiscale adaptive smoothing models for the hemodynamic response function in fMRI, *The annals of applied statistics* **7**(2), 904.

Wasserstein, R. L. and Lazar, N. A. (2016), The ASA Statement on p-Values: Context, Process, and Purpose.

White, T., O'Leary, D., Magnotta, V., Arndt, S., Flaum, M. and Andreasen, N. C. (2001), Anatomic and functional variability: The effects of filter size in group fMRI data analysis, *NeuroImage* **13**(4), 577–588.

Williamson, T., Eliasziw, M. and Fick, G. H. (2013), Log-binomial models: Exploring failed convergence, *Emerging Themes in Epidemiology* **10**(1).

Wongrakpanich, S., Wongrakpanich, A., Melhado, K. and Rangaswami, J. (2018), A Comprehensive Review of Non-Steroidal Anti-Inflammatory Drug Use in The Elderly., *Aging and disease* **9**(1), 143–150.

Worsley, K. and Friston, K. (1995), Analysis of fMRI Time-Series Revisited?–Again, *NeuroImage* **2**(3), 173–181.

Worsley, K. J., Evans, A. C., Marrett, S. and Neelin, P. (1992), A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain, *Journal of Cerebral Blood Flow and Metabolism 12:900-918* .

Worsley, Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J. and Evans, A. C. (1996), A unified statistical approach for determining significant signals in images of cerebral activation, *Human Brain Mapping* **4**(1), 58–73.

Yang, X., Kang, H., Newton, A. T. and Landman, B. A. (2014), Evaluation of statistical inference on empirical resting state fmri, *IEEE Transactions on Biomedical Engineering* **61**(4), 1091–1099.

Yue, Y., Loh, J. M. and Lindquist, M. A. (2010), Adaptive spatial smoothing of fMRI images, Technical report.

Zhou, M., Badre, D. and Kang, H. (2019), Double-wavelet transform for multisubject task-induced functional magnetic resonance imaging data, *Biometrics* **75**(3), 1029–1040.