STATISTICAL METHODS FOR INTEGRATING DISPARATE DATA SOURCES

by

Prosenjit Kundu

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

April, 2020

Abstract

My thesis is about developing statistical methods by integrating disparate data sources with real data applications, and identifying gene-environment interactions (G × E) in more extensive studies using existing analytical methods. We propose a general and novel statistical framework for combining information on multivariate regression parameters across multiple different studies which have varying level of covariate information (Chapter 2). We illustrate the method using real data for developing a breast cancer risk prediction model. We propose a generalized method of moments (GMM) approach for analyzing two-phase studies where we take into account the dependent structure of the datasets across the two-phases (Chapter 3). We illustrate the method using real data on Wilm's tumor, a common type of kidney cancer in children. We analyze the largest gene by smoking interaction study for pancreatic ductal adenocarcinoma risk conducted to date using existing statistical methods (Chapter 4).

Primary Readers

Nilanjan Chatterjee (Advisor)
Professor
Department of Biostatistics & Department of Oncology

Bloomberg School of Public Health, School of Medicine, The Johns Hopkins University

Alison Patricia Klein
Professor
Department of Oncology
School of Medicine, The Johns Hopkins University

Mei-Cheng Wang
Professor
Department of Biostatistics
Bloomberg School of Public Health, The Johns Hopkins University

Debashree Ray
Assisstant Professor
Department of Epidemiology
Bloomberg School of Public Health, The Johns Hopkins University

Alternate Readers

Elizabeth L. Ogburn
Assisstant Professor
Department of Biostatistics
Bloomberg School of Public Health, The Johns Hopkins University

Xiaobin Wang
Professor
Department of Population, Family and Reproductive Health
Bloomberg School of Public Health, The Johns Hopkins University

Acknowledgments

As this long voyage of my PhD comes to an end, I want to take this moment to express my sincere gratitude to all the people who brought me to this significant stage of my life.

I want to thank my Ph.D. advisor, Dr. Nilanjan Chatterjee, for supporting me during the past four years. He is a great teacher and an incredible mentor. One of the greatest things about Nilanjan-Da is the way he thinks. I remember many instances where he would explain the intuition behind a problem in a simple way that left me awestruck. I am genuinely grateful to him for having me injected with the ability to think about a problem intuitively before jumping into the mathematical details. I am thankful to him for helping me enrich my knowledge skills during our regular meetings. I also value his mentorship above and beyond academia, guiding and motivating me during difficult times.

Besides my thesis advisor, I would like to thank Dr. Alison P. Klein, Dr. Mei-Cheng Wang, and Dr. Debashree Ray, for serving on my thesis committee. I want to pay my special regards to Alison for introducing me to the exciting research area in pancreatic cancer and guiding me with the fourth chapter of my thesis. I am grateful to Dr. Elizabeth L. Ogburn and Dr. Xiaobin Wang for serving as alternate members of my thesis committee.

I remain eternally grateful to our department chair, Dr. Karen Bandeen-Roche, for all her support and advice during the past years. I want to take this opportunity to thank Dr. Bimal Kumar Roy (Indian Statistical Institute, Kolkata), Dr. Ciprian Crainiceanu, and Dr. Karen Banden Roche, who first provided me the opportunity to work (as an intern) in the Department of Biostatistics, JHSPH, back in 2013. I am extremely grateful to them. My special thanks to Ciprian and the SMART group for having me introduced to the exciting and emerging area of wearable computing.

The departmental admissions, administration & finance, information systems were an integral part of my PhD journey. I want to acknowledge all of them, and especially, I would like to thank Mary Joy Argo, for sharing her expertise and solving all my academic issues; Maria Beason, for assisting me in scheduling educational meetings; Debbie Cooper, Saima Ahmed, Fallon Bachman, Debra Moffitt, Patty Hubbard for helping me with administrative concerns whenever needed, and Jiong Yang, for assisting me with any cluster-specific issues. A huge thanks to all.

I had the excellent opportunity to work with some great collaborators: Dr. Runlong Tang, Dr. Allison Meisner, Dr. Evelina Mocci, Panc4, and PanScan members, Dr. Raymond J. Carroll, Dr. Joe Coresh, Natalie Daya, Dr. Liz Selvin, Brian Mcgrath, Statistical Genetics group and I am genuinely thankful for all their insights and fruitful discussions. I am also grateful to PCORI and NIH for their financial support.

The charming city, Baltimore, will forever reamain close to my heart. In the past few years, Baltimore has become my hometown and I am lucky to have made amazing friends here. Thank you all for supporting me throughout this

entire process and helped me in maintaining a work-life balance. I want to start by thanking Kayode Sosina for being an awesome officemate and for being just a phone call away for intense research discussions. Special thanks to Parichoy Pal Choudhury, Sumit Sahu, and Shatabdi Pal for making me feel home away from home. They were my guardians, who kept me safe from grief and failures. A heartfelt thank you to Kunal Kundu for all his visits during my stay in Baltimore and for cheering me up at critical times. Thanks to Debangan Dey, Sayan Ghoshal, Soumya Banerjee, Anindya Bhaduri, Arunima Banerjee, and Neha Agarwala. They have become my family now. Thank you for being a part and celebrating the exciting moments of my life with me, be it hanging-out on weekends, studying in the library, planning sudden trips, eating at buffets, and sometimes gambling. My PhD journey has become an easy ride for these people. Some other friends who deserve special mention are Lacey Etzkorn, Linda Gai, Batel Blechter, Haoyu Zhang, Dora Zhang, Subhra Shankar Koley, Amartya Bhattacharjee, Ipshita Bhattacharya, Tushita Mukhopadhyay, Sayantan Dutta, Roshni Roy, Diptavo Dutta among others. To my long list of friends located around the world: Lokaditya Ryali, Avijit Singh, Arnab Das, Debarya Dutta, Prakash Chakraborty, Sandipan Chattopadhyay, Indrayudh Ghoshal, Sourav Sarkar (chutku), Indranil Bhattacharya and others, thank you for all your visits, texts and phone calls.

I want to thank Neha Agarwala, one of the most important persons of my life. Thank you for being there with me and making this journey worthwhile. Thank you for all the beautiful memories, for inspiring me and being a wonderful friend.

I can't thank enough my family: Maa (Soma Kundu), Baba (Debdas Kundu), Dada (Ankit Kundu), and Boudi (Shrutilekha Roy), for all their endless, unparalleled love and support. I couldn't have asked for a better childhood friend than my dada. Maa and baba have always sacrificed their wishes to fulfill my dreams. They took the best care of me and provided the best academic environment at home. Whatever I am today is because of my family. Thank you, baba, maa, dada , and boudi.

Dedications

Dedicated to Maa, Baba, Dada, Boudi, Neha, and friends for all your love and inspiration

Table of Contents

Table of Contents **vi**

List of Tables **x**

List of Figures **xii**

1 Introduction **1**

2 Generalized Meta-Analysis **8**

2.1 Introduction 8

2.2 Models and Methods 10

2.2.1 Formulation of the model 10

2.2.2 A Special Case Involving Linear Regression Model 11

2.2.3 Generalized meta-analysis 13

2.2.4 Iterated Reweighted Least Square Algorithm 18

2.2.5 Diagnostic Test for Model Violation 21

2.3 Simulations 22

2.3.1 Homogeneous Population 23

2.3.2 Heterogeneous Population 28

2.3.3 Power Evaluation of the Diagnostic Test ($T_{GENMETA}$) . . 29
2.4 Real Data Analysis 32
2.5 Discussion 39

3 Analysis of Two-Phase Studies using Generalized Method of Moments 43
3.1 Introduction 43
3.1.1 Model Formulation 46
3.1.2 Sampling Design 46
3.1.3 Method 47
3.2 Simulation Studies Resembling US National Wilms Tumor Data 49
3.3 Application to US National Wilms Tumor Data 53
3.4 Discussion 55

4 Genome-wide Interaction Scan Identifies Gene by Smoking interaction at 2q21.3 for Pancreatic Cancer Risk 61
4.1 Introduction 61
4.2 Results 64
4.2.1 Established GWAS regions 67
4.3 Methods 70
4.3.1 Study sample 70
4.3.2 Genome-wide association genotyping data 71
4.3.3 Smoking and demographic assessment 72
4.3.4 Statistical analyses 72
4.4 Discussion 75

5 Conclusions and Future Work 79

5.0.1 Contribution 79

5.0.2 Future Work 82

References 86

A Chapter 1 127

A.1 Asymptotic Equivalence of GENMETA Estimator and Simple Meta-Analysis Estimator When All the Reduced Models Are the Same to the Maximal Model 127

A.2 Newton-Raphson's Method and Iteratively Reweighted Least Squares Algorithm 130

A.2.1 Case I : ϕ and ϕ_k's are known. 131

A.2.2 Case II : ϕ and ϕ_k's are unknown. 134

A.3 Full proof of theorem 2.2.1 and Checking regularity assumptions in two examples 136

A.3.1 Regulartity Assumptions for Theorem 2.2.1 136

A.4 Simulation Results for Log-normally Distributed Covariates . . . 146

B Chapter 2 148

B.1 Estimating equation in phase-I 148

B.2 Estimating equation in phase-II 150

B.3 Consistency and Asymptotic Normality 150

C Chapter 3 155

List of Tables

2.1 Results on the GENMETA estimators 25

2.2 A Simulation for Understanding Real Data Analysis 27

2.3 Robustness of GENMETA Estimation (Normally Distributed Covariates) 30

2.4 Combined analysis of BCDDP and BPC3 study to develop a multivariate logistic regression model for breast cancer risk. For each cohort within BPC3 and for BCDDP, standard logistic regression model is applied for fitting reduced models including FH (family history), AMEN (age at menarche), AFB (age at first live birth) and WT (weight). Parameter estimates of the reduced models across studies are then combined using standard meta-analysis (meta) or GMeta. For the BCDDP study, a maximal logistic model is fitted including additional covariates mammographic density (MD) and number of previous biopsy (NBIOPS). These estimates are then combined with with estimates of reduced model parameters from BPC3 studies to obtain GMeta estimates of the maximal model. Point estimates (PE) and standard errors (SE) are shown for each analysis. NA means there is no corresponding estimator. The variables analyzed include: FH: binary indicator of family history; AMEN1 and AMEN2: dummy variables associated with age-at-menarche categories ≥ 14, 12–13 and ≤ 11; AFB1 and AFB2: dummy variables associated with age-at-first-live-birth categories ≤ 20, 21–29 and ≥ 30; WT1 and WT2: dummy variables associated with weight categories ≤ 62.6, 62.6–73.1 and ≥ 73.1 in kilograms; NBIOPS: the number of biopsies coded as a conitunous variable and MD: the standardized mammographic density coded as a continuous variable. 38

3.1 Simulation Results (Imitating Real Data Structure) 59

3.2 Wilms Tumor Data: Phase-I strata frequencies and Phase-II sampling design 60

4.1 Region with genome wide significant evidence for SNP by smoking interaction on risk of PDAC 67

A.1 Robustness of GENMETA Estimation (Log-normally Distributed Covariates) 147

C.1 Top-1 SNP (rs1818613) for locus 2q21.3 with association reaching genome wide significance 157

C.2 Odds Ratios and Interaction P-value GWASs loci in Caucasian population. 158

C.2 Odds Ratios and Interaction P-value GWASs loci in Caucasian population (Contd.) 159

C.2 Odds Ratios and Interaction P-value GWASs loci in Caucasian population (Contd.) 160

List of Figures

2.1 Square roots of mean square errors (RMSE) of GENMETA estimators for β_1^*, β_2^* and β_3^* with fixed study sample sizes $n_1 = 300$, $n_2 = 500$ and $n_3 = 1000$ and varying reference sample size n from 10, 30, 50, 70, 100, 200 to 1000. The circle and solid line are for the RMSE's of GENMETA.0; the triangle and dashed line are for those of GENMETA.1; the plus and dotted line are for those of GENMETA.2. 26

2.2 Power curves of simple multivariate meta-analysis test statistic (Q) and $T_{GENMETA}$ for simulated datasets. The long-dashed line is for the simple meta-analysis estimator. The solid and dotted lines are for GENMETA estimators with reference data sample sizes 100 and 500, respectively. Level of the test (α) is set to 0.05. 32

3.1 Mean square errors from real data analysis in case-control design 56

3.2 Mean square errors from real data analysis in balanced design . 57

4.1 Q-Q and Manhattan Plots of Interaction Analysis using Empirical Bayes Approach 65

4.2 Locus plot of 2q21.3 region for the interaction GWAS of pancreatic cancer by smoking using FUMA [161]. a. Extended region of the $TMEM163$ locus that prioritizes genes $TMEM163$ and $CCNT2$. b. Zoomed in regional plot of $TMEM163$ locus with GWAS interaction P-values (SNPs are colored based on r^2), Combined Annotation Dependent Depletion (CADD score), and eQTL P-value. eQTLs are plotted per gene and colored based on tissue types. 66

4.3 Results of colocalization analysis using eCAVIAR and co-loc. SNPs with colocalization probability (CLPP) ≥ 0.001 are shown in this plot. PP.H4 denotes posterior probability of having a common causal snp across eQTL and SNP X Smoking loci. . . . 69

C.1 Q-Q and Manhattan Plots of Interaction Analysis using CML and UML approach 156

C.2 161

Chapter 1

Introduction

In the world of decision making, data is an indispensable ingredient for addressing relevant questions in almost all disciplines of study, including science, humanities, and business. However, in the era of big data, where the collection of data from different studies is increasing in volume, variety, and velocity, it is arduous to share and tackle such massive data across studies. To mitigate this difficulty and to have meaningful results by harnessing knowledge and reasoning from multiple data sets, there is a well-known procedure of harmonizing various data sources in the literature of Big Data, called Data Integration. Formally, data integration is a process of fusing information from multiple, possibly heterogeneous data sources, giving a unified way to draw inference on real-world problems and building generalizability to a larger population [100]. Due to advancements in technology and ease in the availability of modern tools, the research in a variety of fields, including genomic medicine, genetics, clinical trials, epidemiology, and environmental science, has become data-intensive with a deluge of heterogenous data[173, 97, 119, 107, 149]. This has made the researchers to explore and apply data integration strategies to blend all the available information for more discoveries in science, which makes data integration a vital contribution to humanity at present and future.

Disparate data sources are often necessary to answer a scientific question of interest. It is common to see many observational studies with dissimilar, but overlapping, information on some crucial potential risk factors. Here, we provide some real data examples. The Breast Cancer Detection Demonstration Project (BCDDP) Study has information on mammographic density, one of the critical risk factors of breast cancer, apart from other primary risk factors [110]. Since mammograms can be expensive, the study sample size is small. However, the Breast and Prostate Cancer Cohort Consortium, a collection of ten large prospective cohorts, is a more extensive study that has information on the other primary risk factors except mammographic density. In this example, we see two disparate datasets, one a smaller study with a rich set of risk factors and the other a larger study with limited risk factors. Other examples include multiple pediatric cohort studies in the Environmental influences on Child Health Outcomes (ECHO) Program [`https://www.nih.gov/echo`] that aims at understanding the etiological factors affecting children's health outcome by combined analysis of information from existing pediatric cohorts. Disparate risk factors across data can arise from sampling design itself. For example, a two-phase sampling design collects information in two phases where the ascertainment of expensive variables is limited to a judiciously chosen smaller subset of individuals sampled in phase-I. Data across the two phases creates dependent datasets with more variables measured in phase-II [125, 154, 145, 132, 112]. One of the classic examples includes the US National Wilm's Tumor Study that is used by many researchers to simulate a two-phase design [14, 42, 63, 7]. Another example that induces disparate and dependent datasets by design is the UKBiobank, the world's most extensive cohort study to date [25, 50, 49]. Separate substudies within the UKBiobank, including the accelerometry study and imaging study, create disparity on the group of measured variables across them. Variation in quantification of variables across studies can also induce disparity. For example, some studies quantify smoking behavior as never, former and current, while others quantify as pack-years or number of cigarettes smoked per day. All these examples fall under the category of disparate data sources.

Observational studies are designed to generate a hypothesis, where researchers are often interested in the association of a risk factor with an outcome after adjusting for all possible measured confounders under a regression model framework. Stitching together the partial information available across disparate data sources is a potential solution where it is difficult to infer from a single data source due to the unavailability of some of the variables or due to low power. One of the most popular statistical tools, due to logistic convenience and statistical efficiency, used in data integration is meta-analysis. Meta-analysis is a process of synthesizing summary-level information (estimates and standard errors) on common parameters of interest (e.g. log odds ratio, treatment effect) across studies[47, 48, 82]. In large cohorts like UKBiobank(`http://www.ukbiobank.ac.uk`), CKBiobank(`http://www.ckbiobank.org/site/`) where both sample size and the number of covariates measured are huge, model fitting is a daunting computational task. In modern GWAS, clinical trials, and many other epidemiological studies, sharing of data across the studies is a big concern due to various privacy, ethical, and logistical issues. In these troubling situations, meta-analysis plays a significant role by following the divide and conquer approach. It combines the summary-level information across studies, e.g., model parameters, to make conclusions on a scientific question of interest, eventually aiding in decisions for making policies. However, with disparate data sources, we cannot combine the parameter estimates across the data sources as the parameters have different interpretation across studies. For example, it is known that mammographic density and weight are negatively correlated. In this situation, the association of weight with breast cancer after adjusting for mammographic density will have different interpretation with that without adjustment. This motivates us to develop a general statistical framework for integrating disparate data sources.

In chapter 1, we develop a generalized meta-analysis approach for combining information on multivariate regression parameters across multiple different studies which have varying level of covariate information [95]. Using algebraic relationships between regression parameters in various dimensions, we specify a set of moment equations for estimating parameters of a maximal model through the information available from sets of parameter estimates from a series of reduced models available from the different studies [28]. The specification of the equations requires a reference dataset to estimate the joint distribution of the covariates. We propose to solve these equations using the generalized method of moments (GMM) approach, with the optimal weighting of the equations taking into account uncertainty associated with estimates of the parameters of the reduced models [80, 69]. We describe extensions of the iterated reweighted least squares algorithm for fitting generalized linear regression models using the proposed framework. Based on the same moment equations, we also propose a diagnostic test for detecting violations of underlying model assumptions, such as those arising due to heterogeneity in the underlying study populations. Methods are illustrated using extensive simulation studies and a real data example involving the development of a breast cancer risk prediction model using disparate risk factor information from multiple studies.

In chapter 2, we analyze two-phase studies using the GMM approach. Two-phase design can reduce the cost of epidemiological studies by limiting the ascertainment of expensive covariates or/and exposures to an efficiently selected subset (phase-II) of a larger (phase-I) study. Efficient analysis of the resulting dataset combining disparate information from phase-I and phase-II, however, can be complex. Most of the existing methods, including semiparametric maximum-likelihood estimator, require the information in phase-I to be summarized into a fixed number of strata [19, 16, 20]. In this paper, we describe a novel method for analysis of two-phase studies where information from phase-I is summarized by parameters associated with a reduced logistic regression model of the disease outcome on available covariates. We then setup estimating equations for parameters associated with the desired extended logistic regression model based on information on the reduced model parameters from phase-I and complete data available at phase-II after accounting for non-random sampling design at phase-II. We use the generalized method of moments to solve overly identified estimating equations and develop the resulting asymptotic theory for the proposed estimator. Simulation studies show that the use of reduced parametric models can lead to more efficient utilization of phase-I data than summarizing into strata. An application of the proposed method is illustrated using the US

National Wilms Tumor study data.

In chapter 3, we apply standard multivariate meta-analysis to high-dimensional genome-wide association studies of pancreatic cancer. Pancreatic cancer is the seventh leading cause of cancer death worldwide with pancreatic ductal adenocarcinoma (PDAC) being the most common subtype (>90%). Inherited genetic changes and cigarette smoking are established independent risk factors of PDAC. The problem of interest is to identify gene by smoking interactions with pancreatic cancer that can better inform the underlying biological pathway leading to pancreatic cancer. Gene-environment interactions play a significant role in the etiology of cancer risk. There are many existing statistical methods to identify gene-environment interactions [67, 151, 31, 121, 29]. The study of discovering such interactions (gene by smoking in our study) can provide insights into the underlying biological pathways leading to PDAC and thus better inform in making public health strategies for pancreatic cancer prevention. We conducted the largest (till date) genome-wide gene-by-environment interaction analysis of 7,937 PDAC cases and 11,774 controls arising from two data sources, the Pancreatic Cancer Case-Control Consortium (PanC4) and the Pancreatic Cancer Cohort Consortium [93, 32]. After a meta-analysis of these two datasets, we identified a statistically significant interaction by smoking status (never, former, current) of SNPs located on 2q21 (P-value $< 5 \times 10^{-9}$). This region includes rs1818613 and is located intronic to TMEM163 and upstream of CCNT2. Genetic variants in this region are strongly associated (p-value ¡10-8) with differential expression of TMEM163 in several tissues, including heart, pituitary, and whole blood, and differential CCNT2 expression in tibial nerve and lung (p-values $< 10^{-6}$) tissue in the GTEx database. Our work provides evidence of the importance of genetic variation in this region in conjunction with cigarette smoking on PDAC risk.

Chapter 2

Generalized Meta-Analysis

2.1 Introduction

In a variety of domains of applications, including observational epidemiologic studies, clinical trials and modern genome-wide association studies, meta-analysis is widely used to synthesize information on underlying common parameters of interest across multiple studies [45, 46, 83, 88]. The popularity of meta-analysis stems from the fact that it can be performed based only on estimates of model parameters and standard errors, avoiding various logistical, ethical and privacy concerns associated with accessing the individual level data that is required in pooled analysis. Moreover, in many common settings, it can be shown that under reasonable assumptions, meta-analyzed estimates of model parameters are asymptotically as efficient as those from pooled analysis [127, 115, 102]. In fact, meta-analysis approaches are now being used in divide and conquer approaches to big data, even when individual level data are potentially available, because of the daunting computational task of model fitting with extremely large sample sizes [87, 54, 37].

In this chapter, we study the problem of multivariate meta-analysis in the setting of parametric regression modeling of an outcome given a set of covariates. In standard settings, if estimates of multivariate parameters for an underlying common regression model and associated covariances are available across all the studies, then meta-analysis can be performed by taking inverse-variance-covariance weighted average of the vector of regression coefficients [156, 140, 84]. In many applications, a typical problem is that different studies include different, but possibly overlapping, sets of covariates. In a large consortium of epidemiologic studies, for example, some key risk factors will be measured across all the studies. Inevitably, however, there will be potentially important covariates which are measured only in some, but not all the studies. It is also possible that some covariates are measured at a more detailed level or with a finer instrument in some studies compared to others. Disparate sets of covariates across studies render standard meta-analysis to be applicable for the development of models only limited to a core set of variables that are measured in the same fashion across all the studies.

We propose a generalized meta-analysis (GENMETA) approach for building rich models using information on model parameters across studies with disparate covariate information. GENMETA is built upon a fundamental mathematical relationship between parameters of two regression models in different dimensions from our recent study [28]. In the current article, we utilize this mathematical relationship to develop a general framework for combining information on parameters of various models of different dimensions within the generalized method of moments framework [71, 81]. We develop an iterated reweighted least square algorithm allowing stable and speedy computation of estimates. The proposed method requires access to a reference dataset for estimating of joint distribution of the covariates in a nonparametric fashion. We show how the reference dataset can be used to derive an optimal estimator and associated variance-covariances even when entire variance-covariance matrices for model parameter estimates may not be obtainable from individual studies.

2.2 Models and Methods

2.2.1 Formulation of the model

Suppose we have parameter estimates $\hat{\theta}_k$ and associated estimates of their covariance matrices S_k from K independent studies which have fitted reduced regression models, for which the likelihood is of the form $g_k(Y \mid X_{A_k}; \theta_k)$, where Y is a common underlying outcome of interest but the vector of covariates X_{A_k} is potentially distinct across the studies. Let X be the set of covariates used across all studies and we assume the true distribution of Y given X can be specified by a maximal regression model $f(Y \mid X; \beta)$. Our goal is to estimate and make inference about β^*, the true value of β, based on summary-level information, $(\hat{\theta}_k, S_k)$ from the K studies.

In the proposed setup, it is possible but not necessary, that one or more of the studies have information on all covariates to fit the maximal model by themselves. Under certain study designs, such as the multi-phase designs [15, 17, 162, 147] and the partial questionnaire design [157], data could be partitioned into independent sets where the maximal model can be fitted on some sets and various reduced models can be fitted on others. The maximal model $f(Y \mid X; \beta)$ and the reduced models $g_k(Y \mid X_{A_k}; \theta_k)$ may have different parametric forms, such as logistic and probit models when Y is a binary disease outcome. This setup also allows incorporation of covariates which may be measured more accurately or in a more refined fashion in some studies than others. For example, different studies may include two types of measurements, namely, Z_1 and Z_2, for the same covariate, with Z_2 a more refined measurement. In this case the different reduced models may include Z_1 or Z_2, but we require that the reference dataset includes both Z_1 and Z_2. In the maximal model, we can enforce that Y is independent of Z_1 given Z_2 by setting the regression parameters associated with Z_1 to be zero.

If all of reduced models were the same, i.e. all studies have the same covariate information, we have $X_k = X$, $\theta_k = \beta$ and $g_k = f$ for each k, and the common parameter of interest β^* can be efficiently estimated by the fixed-effect meta-analysis estimator $\hat{\beta}_{\text{meta}} = \sum_{k=1}^{K} (\sum_{k=1}^{K} S_k^{-1})^{-1} S_k^{-1} \hat{\theta}_k$, the variance of which, in turn, can be estimated by $\hat{\Sigma}_{\text{meta}} = (\sum_{k=1}^{K} S_k^{-1})^{-1}$ [156, 140, 84].

2.2.2 A Special Case Involving Linear Regression Model

As readers may find it counter-intuitive to comprehend how it is possible to estimate parameters of the maximal model as no single study may have ascertained Y and all components of X simultaneously, following we give a linear model example to help develop insight into the problem. Suppose, one is interested in developing a multiple linear regression model for Y based on a set of covariates

X in the form

$$Y = \alpha + \sum_{k=1}^{K} \beta_k X_k + \epsilon$$

where it is further assumed that $\epsilon \sim N(0, \sigma^2)$. Without loss of generality, we will assume that all the variables Y, $X_1, \ldots, X_K$ are standardized to have mean zero and variance one. Under this model, the population parameter $\beta = (\beta_1, \ldots, \beta_K)^T$ can be expressed as $\beta = E(X^T X)^{-1} E(X^T Y) = R^{-1} E(X^T Y)$, where R is the population correlation matrix of X. Now, suppose we have no data available on Y and mutivariate X on the same sample, but we have estimates available for parameters (θ_k, $k = 1, \ldots K$) for univariate linear regression models of the form

$$Y = \theta_k X_k + \psi_k.$$

From above $\theta_k = E(X_k Y)$ and thus $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$, provides an estimate of the cross product terms, $E(X^T Y)$, which is required in estimating β. Further, if we have a reference dataset, which has information on multivariate X but is not required to be linked to Y, it can be use to estimate R, as $\hat{R}$ say, and a consistent estimate of β can be obtained simply as $\hat{\beta} = \hat{R}^{-1}\hat{\theta}$. Thus, this simple derivation shows that it is possible to estimate parameters of a multiple regression model using information on parameters of a series of univariate regression models and a reference dataset. In fact, this observation that information on univariate regression parameters (known as summary-level statistics) can be utilized to reconstruct estimates of parameters of multivariate regression model has revolutionized the field of statistical genetics. Recently, a large variety of methods have been developed for the inference on parameters underlying multivariate regression models utilizing widely available summary-level results from large

GWAS and reference datasets to estimate linkage disequlibrium across genetic markers [168, 23, 173, 128]. In the following, we show a more general statistical formulation of the problem that allows consideration of non-linear models and use of information from arbitrary types of reduced models as opposed to simply univariate models.

2.2.3 Generalized meta-analysis

To understand the approach, we start with a simple example. Suppose, we have two studies, $K = 2$. In the first study, let us assume we have measured values on two covariates, say, X_1 and X_2, while in the second study we have measured values on a different set, but overlapping, of covariates, say, X_2 and X_3. Also, both the studies have measured values on the outcome of interest, say, Y which we assume to be a binary variable. We assume both the studies are independent, employ a random sampling design and the same probability law of (Y, X_1, X_2, X_3) holds in all the underlying populations. Someone fits a logistic regression model, g, in each of the two studies and obtain the mle's of the respective model parameters and their standard errors, denoted by $(\hat{\boldsymbol{\theta}}_{A_k}, \boldsymbol{S}_{A_k})$, $k = 1, 2$. To be specific, say, we fit a model to data from the first study, ssuming that $Y|X_1, X_2 \sim Bernoulli(\{1+exp(-(\theta_{10}+\theta_{11}X_1+\theta_{12}X_2))\}^{-1})$, using the standard glm package in R. W estimate the maximum-likelihood estimator of the parameter vector, $(\hat{\theta}_{10}, \hat{\theta}_{11}, \hat{\theta}_{12})^T$ and denote it by $\hat{\boldsymbol{\theta}}_{A_1}$, where $A_1 = \{1, 2\}$. And, we denote its standard error by $\boldsymbol{S}_{A_1}$. Similarly, we fit a logistic regression model to the other data from the other study to obtain the estimates, $(\hat{\boldsymbol{\theta}}_{A_2}, \boldsymbol{S}_{A_2})$. Now, suppose we are provided with the summary-level information,

$((\hat{\boldsymbol{\theta}}_{A_1}, \boldsymbol{S}_{A_1}), (\hat{\boldsymbol{\theta}}_{A_2}, \boldsymbol{S}_{A_2}))$, only. Let $\boldsymbol{X} = (X_1, X_2, X_3)$ denote the full set of covariates across studies. Also, assume that the true maximal model, f, of $Y|\boldsymbol{X}$ is $Y|\boldsymbol{X} \sim Bernoulli(\{1 + exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2))\}^{-1})$. We aim for estimating the true value, $\boldsymbol{\beta}^*$, of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$.

Let $\boldsymbol{S}_k(y|\boldsymbol{x}_{A_k}; \boldsymbol{\theta}_{A_k}) = \frac{\partial \log g(y|\boldsymbol{x}_{A_k}; \boldsymbol{\theta}_{A_k})}{\partial \boldsymbol{\theta}_{A_k}} = (y - expit(\boldsymbol{\theta}_{A_k}^T \boldsymbol{x}_{A_k}))\boldsymbol{x}_{A_k}$ denote the score function in study $k = 1, 2$, solving which we get the mle's, $(\hat{\boldsymbol{\theta}}_{A_1}, \hat{\boldsymbol{\theta}}_{A_2})$. Irrespective of the correct or incorrect specification of the reduced models fitted to the studies, $E_{P^*}\boldsymbol{S}_k(y|\boldsymbol{x}_{A_k}; \boldsymbol{\theta}_{A_k}) = 0$ holds true, where, P^* is the true probability law. Writing the expectation as an iterated expectation , we get

$$\begin{aligned} E_{Y,\boldsymbol{X}}\boldsymbol{S}_k(Y|\boldsymbol{X}_{A_k}; \boldsymbol{\theta}_{A_k}) &= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}[(Y - expit(\boldsymbol{\theta}_{A_k}^T \boldsymbol{X}_{A_k}))\boldsymbol{X}_{A_k}] \\ &= E_{\boldsymbol{X}}[(expit(\boldsymbol{\beta}^T \boldsymbol{X}) - expit(\boldsymbol{\theta}_{A_k}^T \boldsymbol{X}_{A_k}))\boldsymbol{X}_{A_k}] \end{aligned}$$

Denote $(expit(\boldsymbol{\beta}^T \boldsymbol{X}) - expit(\boldsymbol{\theta}_{A_k}^T \boldsymbol{X}_{A_k}))\boldsymbol{X}_{A_k}$ by $\boldsymbol{U}_k(\boldsymbol{X}_{A_k}, \boldsymbol{X}; \boldsymbol{\theta}_{A_k}, \boldsymbol{\beta})$. Then, we have the following key equation that connects the full model parameter, $\boldsymbol{\beta}$ and reduced model parameters $\boldsymbol{\theta}_{A_k}$, for $k = 1, 2$.

$$E_{\boldsymbol{X}}\boldsymbol{U}_k(\boldsymbol{X}_{A_k}, \boldsymbol{X}; \boldsymbol{\theta}_{A_k}, \boldsymbol{\beta})|_{(\boldsymbol{\beta}=\boldsymbol{\beta}^*, \boldsymbol{\theta}=\boldsymbol{\theta}^*)} = 0 \tag{2.1}$$

where $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$ denote the true values.

Reference dataset to estimate joint distribution of all risk-factors : To evaluate the above expectation, we need a reference data set on all the covariates, X_1, X_2, X_3 to empirically estimate the distribution function $F(\boldsymbol{X})$.

Since, we may not have individual level information from the studies, we assume there is a reference data set independent of the studies. Later, we show through simulation studies that the sample size for the reference data set need not be large to reach the plateau of efficiency. Replacing the study parameters by their estimates from the summary-level information and the expection with its sample version, we get a sample version of the LHS in eqn(1), $\frac{1}{n_{ref}}\sum_{i=1}^{n_{ref}}\boldsymbol{U}_k(\boldsymbol{X}_{A_k,i},\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}_{A_k},\boldsymbol{\beta})$, denote it by $\boldsymbol{U}_{nk}(\boldsymbol{\beta})$ for $k = 1, 2$, where n_{ref} is the sample size of the chosen reference data set. Thus, from each study provides a single estimation equation of the form (1) which in turn gives a sample vector, like $\boldsymbol{U}_{nk}(\boldsymbol{\beta})$ for kth study.

Now, we stack those vectors from two studies into a single vector, denoted by, $\boldsymbol{U}_n(\boldsymbol{\beta}) = (\boldsymbol{U}_{n1}^T(\boldsymbol{\beta}), \boldsymbol{U}_{n2}^T(\boldsymbol{\beta}))^T$. Our goal, now, boils down to solving $\boldsymbol{U}_n(\boldsymbol{\beta}) = \mathbf{0}$ for $\boldsymbol{\beta}$. We might not be able to exactly solve this equation as the number of equations is, $dim(\boldsymbol{\theta}_{A_1}) + dim(\boldsymbol{\theta}_{A_2}) = 6$, is greater than the dimension of $\boldsymbol{\beta} = 4$. We try to find a solution close to zero, where the concept of generalised method of moments(GMM) perfectly fits in by minimizing the quadratic form $Q_n(\boldsymbol{\beta}) = \boldsymbol{U}_n^T(\boldsymbol{\beta})\boldsymbol{C}\boldsymbol{U}_n(\boldsymbol{\beta})$ for a positive-definite matrix, $\boldsymbol{C}$. We define our GMeta estimator to be

$$\hat{\boldsymbol{\beta}}_{GMeta} := \text{argmin}_{\boldsymbol{\beta}}\, Q_n(\boldsymbol{\beta}).$$

Similary, we extend the approach described above to K independent studies, with f and g_k's belonging to the class of generalized linear model. Without any loss of generality, we assume the functional form of the reduced models to be same across studies, i.e., $g_k = g$ for $k = 1, \ldots, K$.

Asymptotics of $\hat{\boldsymbol{\beta}}_{GMeta}$: Assume the study summary statistics $\hat{\boldsymbol{\theta}}_k$'s are independent; ${n_k}^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \rightarrow N(0, \boldsymbol{\Sigma}_k)$ in distribution and $\lim_{n\rightarrow\infty} n_k/n = c_k > 0$ for each k; and the reference sample is independent of the study samples. Denote $\boldsymbol{\Gamma} = E(\partial U(\boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}^\star)/\partial \boldsymbol{\beta} \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^*})$, $\boldsymbol{\Delta} = E(U(X; \boldsymbol{\beta}^*, \boldsymbol{\theta}^\star)U^T(X; \boldsymbol{\beta}^*, \boldsymbol{\theta}^\star))$ and $\boldsymbol{\Lambda} = diag(\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K)$, where $\boldsymbol{\Lambda}_k = (1/c_k)W_k\boldsymbol{\Sigma}_k W_k^T$ and $W_k = E\partial u_k(X; \boldsymbol{\beta}^*, \boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k \mid_{\boldsymbol{\theta}_k=\boldsymbol{\theta}_k^*}$ for each k.

Theorem 2.2.1 (Consistency and Asymptotic Normality of $\hat{\boldsymbol{\beta}}$)**.** *Suppose the positive semi-definite weighting matrix $\hat{C} \rightarrow C$ in probability. Then, under Assumptions (A1)-(A4) in the appendix, $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}^*$ in probability. Further, given β^* is an interior point and under additional Assumptions (A5)-(A9) in the appendix, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ converges in distribution to the normal distribution $N(0, (\boldsymbol{\Gamma}^T C \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T C (\boldsymbol{\Delta} + \boldsymbol{\Lambda}) C \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T C \boldsymbol{\Gamma})^{-1})$.*

The optimal C that minimizes the above asymptotic covariance matrix is $C_{\text{opt}} = (\Delta + \Lambda)^{-1}$ and the corresponding optimal asymptotic covariance matrix is $\{\Gamma^T(\Delta + \Lambda)^{-1}\Gamma\}^{-1}$. Because C_{opt} itself depends on unknown underlying parameters, it requires iterative evaluation. In our applications, we first evaluate an initial GENMETA estimator with a simple choice of $\hat{C}$ such as the identity matrix. We then obtain the iterated GENMETA estimator by continuing to set $\hat{C} = \hat{C}_{\text{opt}}$ based on the latest parameter estimate till convergence. By Theorem 2.2.1, $\hat{\beta}$ with C_{opt} approximately follows a Gaussian distribution with mean β^* and covariance matrix

$$[\Gamma^T\{\frac{1}{n}\Delta + \text{diag}(\frac{1}{n_1}W_1\Sigma_1 W_1^T, \ldots, \frac{1}{n_K}W_K\Sigma_K W_K^T)^{-1}\}\Gamma]^{-1}, \tag{2.2}$$

which indicates that the precision of GENMETA depends on the size of the reference sample, n, as well as on those of the studies, n_k. However, as we will see in Section 3, the study sample sizes are the dominating factor controlling the precision of GENMETA and with fixed n_k's, the precision of GENMETA quickly reaches plateau as a function of n.

For the implementation of the optimal GENMETA and the variance estimation of any of the GENMETA estimators, one needs to have valid estimates of Λ_k, which depend on Σ_k, the asymptotic covariance matrices of the estimates of the reduced model parameters. Ideally, the studies should provide robust estimates of the covariance matrices, such as the sandwich covariance estimators, so that they are valid irrespective of whether the underlying reduced models are correctly specified or not. In practice, however, while some kind of estimates of standard errors of the individual parameters are expected to be available from a study, obtaining the desired robust estimate of the entire covariance matrix could be difficult. When no estimate of Σ_k is available from the kth study, one can take the advantage of the reference sample to estimate it by $\hat{\Sigma}_k^{\text{ref}} = \hat{J}^{-1}\hat{V}\hat{J}^{-1}$, where $\hat{J} = P_n[E_{Y|X}\{\nabla_{\theta_k} s_k(\theta_k)\}]|_{\theta_k=\hat{\theta}_k}$, $\hat{V} = P_n[E_{Y|X}\{s_k(\theta_k)s_k(\theta_k)^T\}]|_{\theta_k=\hat{\theta}_k}$, $s_k(\hat{\theta}_k) = s_k(Y \mid X_{A_k}; \theta_k)|_{\theta_k=\hat{\theta}_k}$, $\hat{\theta}_k$ is a consistent estimator of $\theta_k^\star$, $\hat{E}_{Y|X}$ is the expectation with respect to the distribution of $Y \mid X$ with $\beta^\star$ replaced by a consistent estimator $\hat{\beta}$, and P_n is the empirical measure with respect to the reference sample. Further, assuming $E_{Y|X}\{\nabla_{\theta_k} s_k(\theta_k)\}|_{\theta_k=\theta_k^*} = \nabla_{\theta_k} E_{Y|X}\{s_k(\theta_k)\}|_{\theta_k=\theta_k^*}$, it follows
$\Lambda_k = (1/c_k)E_{(Y,X)}\{s_k(\theta_k)s_k(\theta_k)^T\}|_{\theta_k=\theta_k^*}$, which can be estimated by $\hat{\Lambda}_k^{\text{ref}} = (1/c_k)P_n[E_{Y|X}\{s_k(\theta_k)s_k(\theta_k)^T\}]|_{\theta_k=\hat{\theta}_k}$. For example, suppose $Y \mid X$ and $Y \mid X_{A_k}$ follow logistic distributions with parameters $\beta^\star$ and θ_k, respectively. Denote

$X = (1, X^T)^T$ and $X_{A_k} = (1, X_{A_k}^T)^T$. Then,

$$\hat{\Lambda}_k^{\text{ref}} = \frac{1}{c_k} P_n[\{(1+e^{X_{A_k}^T \hat{\theta}_k})^{-2}(1+e^{-X^T\beta})^{-1}+(1+e^{-X_{A_k}^T \hat{\theta}_k})^{-2}(1+e^{X^T\hat{\beta}})^{-1}\}X_{A_k}X_{A_k}^T]. \tag{2.3}$$

In section 2.3, we will study the properties of the GENMETA estimators using either covariance matrices estimated from studies or the reference sample.

It is insightful to explore the connection between GENMETA and standard meta-analysis when all of the reduced models are identical to the maximal model, that is, when $\theta_k^* = \beta^*$, $X_{A_k} = X$ and $g_k = f$ for each k. Under this setup, the moment vector evaluated at the true parameters becomes zero for each study, i.e. $u_k(X; \beta^*, \theta_k^*) = u_k(X; \beta^*, \beta^*) = 0$. This simplification implies $\Delta = 0$ and thus the optimal weighting matrix is $C_{\text{opt}} = \Lambda^{-1} = \text{diag}(c_1\Sigma, \ldots, c_K\Sigma)$, where Σ is the inverse of the Fisher's information matrix of f. Denote by $\hat{\beta}_{\text{opt}}$ the GENMETA estimator with a consistent estimator of C_{opt}. Then, by arguments similar to those in the proof of Theorem 2.2.1, $\hat{\beta}_{\text{opt}}$ can be expressed as

$$\hat{\beta}_{\text{opt}} = \hat{\beta}_{\text{meta}} + o_p(1/n^{1/2}),$$

which implies that $\hat{\beta}_{\text{opt}}$ and $\hat{\beta}_{\text{meta}}$ are asymptotically equivalent in terms of limiting distributions.

2.2.4 Iterated Reweighted Least Square Algorithm

GENMETA computation involves minimization of a quadratic form, $Q_C(\beta) = U_n^T(\beta, \hat{\theta})CU_n(\beta, \hat{\theta})$, with a known weighting matrix C. Next, we derive the iterated reweighted least squares algorithm for minimizing the quadratic form, assuming that the maximal and reduced models belong to the class of generalized linear models [117]. Specifically, the densities of $Y \mid X$ and $Y \mid X_{A_k}$ are of the forms $\exp(\{1/a(\phi)\}(yh(x^T\beta^\star) - b\{h(x^T\beta^\star)\}) + c(y;\phi))$ and $\exp(\{1/a(\phi_k)\}(yh(x_{A_k}^T\theta_k) - b\{h(x_{A_k}^T\theta_k\}) + c(y;\phi_k))$, respectively, where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, $h(\cdot) = b'^{-1}\{g^{-1}(\cdot)\}$, g is a monotone and differentiable link function, and ϕ and ϕ_k are the dispersion parameters of the maximal and the kth reduced models, respectively.

First, we assume that the dispersion parameters, ϕ and ϕ_k's, are known and later we will relax this assumption. For this case, it follows, for each k,

$$u_k(x;\beta,\theta_k) = r_k(x;\beta,\theta_k,\phi_k)x_{A_k}, \tag{2.4}$$

where $r_k(x;\beta,\theta_k,\phi_k) = \{1/a(\phi_k)\}(g^{-1}(x^T\beta) - g^{-1}(x_{A_k}^T\theta_k))h'(x_{A_k}^T\theta_k)$. Then, the empirical moment vector is $U_n(\beta,\hat{\theta}) = P_n(u_1(X;\beta,\hat{\theta}_1)^T,\ldots,u_K(X;\beta,\hat{\theta}_K)^T)^T$. The Newton-Raphson (NR) method for searching the minimizer of $Q_C(\beta)$ can be written as

$$\beta^{(t+1)} = \beta^{(t)} - (X_{\text{rbind}}^T W^* X_{\text{rbind}})^{-1} X_{\text{rbind}}^T W X_{A_{\text{diag}}} C X_{A_{\text{diag}}}^T r \tag{2.5}$$

where $X_{\text{rbind}} = 1 \otimes X$ and $X_{(n\times p)}$ is the reference data matrix; $X_{A_{\text{diag}}} = \text{diag}(X_{A_1},\ldots,X_{A_K})$ and $X_{A_k(n\times d_k)}$ is the reference data matrix for the kth study; $W = \text{diag}(W_1,\ldots,W_K)$, $W_k = \text{diag}(w_{k1},\ldots,w_{kn})$ and
$w_{ki} = (1/(a(\phi_k)g'\{g^{-1}(X_i^T\beta^{(t)}))\})h'(X_{A_k,i}^T\hat{\theta}_k)$ for $k = 1,\ldots,K$; $i = 1,\ldots,n$; W^* is the sum of $WX_{A_{diag}}CX_{A_{diag}}^T W$ and $\text{diag}(r^T X_{A_{diag}} C X_{A_{diag}}^T L)$, a diagonalized matrix from a vector; $r = (r_1^T,\ldots,r_K^T)^T$, $r_k = (r_{k1},\ldots,r_{kn})^T$ and

$r_{ki} = r_k(X_i; \beta^{(t)}, \hat{\theta}_k, \phi_k)$; and $L = \mathrm{diag}(L_1, \ldots, L_K)$, $L_k = \mathrm{diag}(l_{k1}, \ldots, l_{kn})$ and

$l_{ki} = -g''\{g^{-1}(X_i^T\beta^{(t)})\}/(a(\phi_k)[g'\{g^{-1}(X_i^T\beta^{(t)})\}]^3 h'(X_{A_k,i}^T\hat{\theta}_k))$. Equation (2.5) implies that the Newton-Raphson's method is an iterated reweighted least squares algorithm.

When ϕ and ϕ_k's are unknown, we propose to first obtain the GENMETA estimator $\hat{\beta}$ of $\beta^\star$ as above with $\phi_k's$ replaced by $\hat{\phi}_k$'s. Next, we consider the estimation of $\phi^\star$, the true value of ϕ. For the kth reduced model, we have an additional score function with respect to ϕ_k, from which, similar to equation (2.4), we can obtain

$$u_k(X; \beta, \phi, \theta_k, \phi_k) = -\frac{a'(\phi_k)}{a^2(\phi_k)}(g^{-1}(X^T\beta)h(X_{A_k}^T\theta_k) - b\{h(X_{A_k}^T\theta_k)\}) + q_k(X; \beta, \phi, \phi_k),$$

where $q_k = E_{Y|X}\{c'(Y; \phi_k)\}$ and $c'(Y; \phi_k)$ is the derivative of $c(Y; \phi_k)$ with respect to ϕ_k. Then, the empirical moment vector for ϕ is $U_n(\phi) = P_n(u_1(X; \hat{\beta}, \phi, \hat{\theta}_1, \hat{\phi}_1)^T, \ldots, u_K(X; \hat{\beta}, \phi, \hat{\theta}_K, \hat{\phi}_K)^T)^T$. To estimate $\phi^\star$, we need to compute the minimizer of $U_n(\phi)^T C U_n(\phi)$, where C is a known weighting matrix. The Newton-Raphson steps can be written as

$$\phi^{(t+1)} = \phi^{(t)} - J_n^{-1}(\phi^{(t)})D_n(\phi^{(t)}), \tag{2.6}$$

where $J_n(\phi) = U_n^T(\phi)Cd^2q_n(\phi)/d\phi^2 + (dq_n(\phi)/d\phi)^T C dq_n(\phi)/d\phi$, $D_n(\phi) = U_n^T(\phi^{(t)})Cdq_n(\phi)/d\phi$ and $q_n(\phi) = P_n(q_1(X; \hat{\beta}, \phi, \hat{\phi}_1), \ldots, q_K(X; \hat{\beta}, \phi, \hat{\phi}_K))^T$. In brief, when ϕ and ϕ_k, $k = 1, \ldots, K$, are unknown, we first choose initial estimates $\beta^{(0)}$ and $\phi^{(0)}$. Then, we get the GENMETA estimator $\hat{\beta}$ by using equation (2.5) until a stopping rule is reached. Subsequently, $\phi^{(0)}$, $\hat{\beta}$ and the study estimates are plugged in equation (A.12) and the process is repeated until a stopping rule is reached to get the GENMETA estimator of ϕ^*. In each NR step, the weighting matrix C is estimated by the estimates from the previous step. A software implementing the IRWLS algorithm in the form of an R Package (GENMETA) is available on CRAN and Github repositories through the links `https://cran.r-project.org/package=GENMETA` and `https://github.com/28pro92/packages-GENMETA`, repectively.

2.2.5 Diagnostic Test for Model Violation

GENMETA relies on several modeling assumptions, including homogeneity of the underlying populations with respect to the distribution of covariates and regression parameters, and correct specification of the maximal model. In the absence of individual level data from the different studies, these assumptions could not be tested in the usual manner using traditional diagnostic tests. However, even with summary-level data, some diagnostic testing is possible. In particular, from an intuitive perspective, departure of the GENMETA estimating equations, when evaluated at estimated parameter values, from their expected null value will be indicative of disagreement between the model and the observed data, i.e. the estimates of the parameters from the reduced models from different studies. For example, if the regression parameters underlying the maximal model are highly heterogeneous across studies, then the assumption of a common β in GENMETA will not be able to explain the heterogeneity that is expected to be present in overlapping reduced model parameters across the studies. Specifically, we propose to use the score test based on the statistic, $T_{GENMETA} = nQ_{\hat{C}_{opt}}(\hat{\beta})$, where $\hat{\beta}$ is the GENMETA estimate. When all the underlying assumptions are correct, from the standard generalized method of moments theory, $T_{GENMETA}$ converges in distribution to a χ^2 distribution with $d-p$ degrees of freedom, where d is the total number of GENMETA equations and p is the total number of underlying parameters that are being estimated. The test is only applicable when $d > p$, which arises when different studies have overlapping covariates.

2.3 Simulations

We study the performance of the GENMETA estimators through simulation studies in both idealized and non-idealized settings. In all simulations, we assume that the relationship between a binary outcome variable Y and three covariates (X_1, X_2, X_3) can be described by a logistic regression model of the form

$$Y \mid (X_1, X_2, X_3) \sim \text{Bernoulli}([1+\exp\{-(\beta_0^* + \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* X_3)\}]^{-1}) \quad (2.7)$$

where (X_1, X_2, X_3) follows a multivariate normal distribution with mean $\mu = (\mu_1, \mu_2, \mu_3)$, variance $\sigma^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ and underlying correlations $\rho = (\rho_{12}, \rho_{13}, \rho_{23})$. We chose $\beta_1^* = \beta_2^* = \beta_3^* = \log 1.3$ to reflect a moderate degree of association of the outcome with each covariate after adjusting for the others. We assume existence of three separate studies, where each study fits a reduced logistic model for the outcome Y on two of the covariates in the form

$$Y \mid (X_i, X_j) \sim \text{Bernoulli}((1 + \exp\{-(\theta^*_{0,ij} + \theta^*_{i,ij} X_i + \theta^*_{j,ij} X_j,)\})^{-1}), \qquad (2.8)$$

with X_1 and X_2 included in Study-I, X_2 and X_3 in Study-II and X_1 and X_3 in Study-III. Here, as data for each study are generated using the maximal model, the reduced models are by definition incompatible due to non-collapsibility of the logistic model. We fix the sample size of the studies at $n_1 = 300$, $n_2 = 500$ and $n_3 = 1000$ and vary the sample size of the reference dataset.

2.3.1 Homogeneous Population

We assume that the studies are conducted in the same underlying population from which the reference sample is drawn. Under this setting, there exists a common mean vector $\mu_b = (0, 0, 0)$, common variance vector $\sigma^2_b = (1, 1, 1)$ and common correlation vector $\rho_b = 0.3, 0.6, 0.1)$, that describes the joint distribution of the three covariates across all the underlying populations. In the first set of simulation, we assume a fixed sample size $n = 50$ for the reference dataset. In all settings, we simulate data (Y, X_1, X_2, X_3) for the underlying studies based on the data generating models as described above and fit the respective reduced models to obtain estimates of the reduced model parameters. For each set of simulated data, we obtain estimates of covariance matrices of the reduced model parameters using robust sandwich estimators based on either the study datasets themselves, or the reference dataset (see (2.3)). We consider three GENMETA estimators: GENMETA.0, which is the initial GENMETA estimator with identity weighting matrix and GENMETA.1 and GENMETA.2, that use covariance estimates from the reference dataset and the studies, respectively.

From the results shown in Table 2.1, we observe that all three GENMETA estimators are nearly unbiased. The standard error estimates, irrespective of whether $\Sigma_k, k = 1, 2, 3$ were estimated using the study data sets or the reference sample, accurately reflected the true standard errors of the GENMETA parameter estimates across different simulations. As a result, the 95% confidence intervals maintained the coverage probability at the nominal level. Among the three GENMETA estimators considered, clearly GENMETA.0, which use the non-optimal choice of $C = I$, is less efficient than GENMETA.1 and GENMETA.2, which, between themselves, had comparable efficiency.

In the same setting as above, when we vary n from 10 up to the maximum of 1000 (Figure 2.1), we observe that the precision of the GENMETA estimates do not increase with n once it reaches a threshold around 100, which is one third of the minimum of the study sample sizes($n_1 = 300$). These thresholds were even smaller for estimation of coefficients associated with X_2, which had weak to moderate correlation with the other covariates in the model. The fact that the reference dataset can be substantially smaller than the study datasets without having much impact on the precision of the GENMETA estimator is encouraging given that accessing reference dataset of large sample size may be difficult in practice.

Finally, we conduct additional simulation studies to obtain more insight into results from the real data analysis (Section 2.4). Here, the settings are similar to before except we assume there are only two studies: study-I fits the maximal logistic regression model involving all the three covariates and study-II involves only two covariates, namely X_1 and X_2. We assume $\rho_I = \rho_{II} = \rho_b$. In our

Table 2.1: Results on the GENMETA estimators

$n = 50$	β_i^*	Bias	SD (ESD$_1$, ESD$_2$)	RMSE	CR	AL
GENMETA.0	β_1^*	.010	.161 (.161, .162)	.161	.968, .964	.642, .636
	β_2^*	.005	.110 (.111, .108)	.110	.958, .960	.434, .423
	β_3^*	-.001	.138 (.143, .142)	.138	.963, .964	.559, .556
GENMETA.1	β_1^*	.005	.117 (.116, .110)	.117	.976, .966	.455, .433
	β_2^*	-.003	.101 (.105, .099)	.101	.964, .955	.411, .386
	β_3^*	.001	.099 (.102, .097)	.099	.973, .961	.402, .381
GENMETA.2	β_1^*	.007	.115 (.116, .111)	.115	.971, .964	.455, .435
	β_2^*	-.003	.102 (.105, .099)	.102	.960, .959	.413, .388
	β_3^*	.003	.098 (.103, .098)	.098	.957, .957	.403, .383

Biases, standard deviation (SD), estimated standard deviation (ESD), square roots of mean square errors (RMSE), coverage rates (CR) and average lengths (AL) of 95% confidence intervals for GENMETA.0 (the initial GENMETA estimator with identity weighting matrix), GENMETA.1 and GENMETA.2 (the iterated GENMETA estimators without and with using the study covariance estimators) in the logistic regression setting. Standard deviations were estimated either using the reference sample (ESD$_1$) or using the covariance estimates of reduced model parameters from the studies (ESD$_2$). Estimated standard deviations are reported by taking averages over simulated datasets. Both estimated SE's are used to construct 95% confidence intervals and their CR's and AL's are reported.

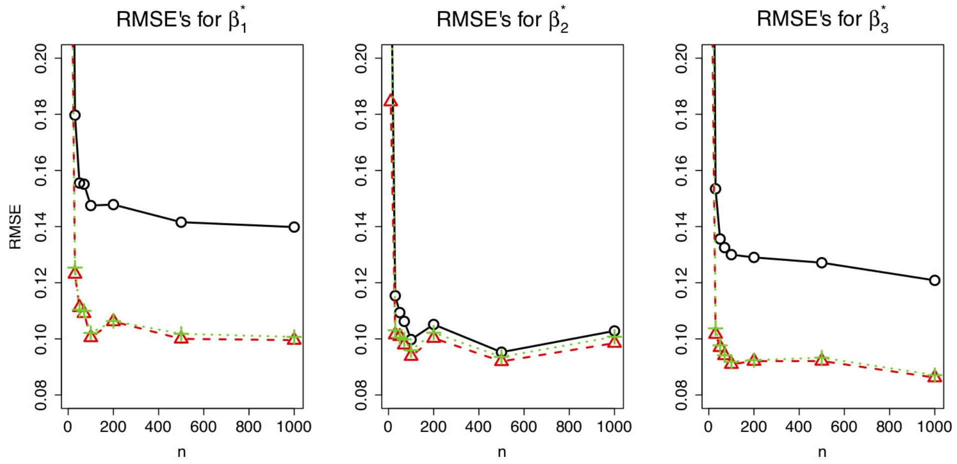


Figure 2.1: Square roots of mean square errors (RMSE) of GENMETA estimators for β^*_1, β^*_2 and β^*_3 with fixed study sample sizes $n_1 = 300$, $n_2 = 500$ and $n_3 = 1000$ and varying reference sample size n from 10, 30, 50, 70, 100, 200 to 1000. The circle and solid line are for the RMSE's of GENMETA.0; the triangle and dashed line are for those of GENMETA.1; the plus and dotted line are for those of GENMETA.2.

estimation, we further considered an added complexity to account for study specific intercept terms for the maximal logistic regression model

$$Y \mid (X_1, X_2, X_3, \text{study}) \sim \text{Bernoulli}([1+\exp\{-(\beta^*_{0,\text{study}}+\beta^*_1 X_1+\beta^*_2 X_2+\beta^*_3 X_3)\}]^{-1})$$

so that the prevalence of the outcome, $\text{pr}(Y = 1)$, could be different across the two studies. In this setting, the maximal set of parameters that are to be estimated through GENMETA can be defined as $\beta^* = (\beta_{0,\text{study-I}}, \beta_{0,\text{study-II}}, \beta_1, \beta_2, \beta_3)$. We simulated data using values of intercept parameters that are identical across the two models, but for estimation we allowed the intercept parameters to be different. For the sake of comparison, we also fitted a reduced model for study-I and conducted a standard multivariate meta-analysis of the underlying common

Table 2.2: A Simulation for Understanding Real Data Analysis

	Study I		Study II	Meta	GENMETA	
	Maximal	Reduced	Reduced	Reduced	Reduced	Maximal
β_i^*	PE (SD)	PE (SD)	PE (SD)	PE (SD)	PE (SD)	PE (SD)
β_1^*	.270 (.149)	.429 (.116)	.424 (.037)	.424 (.035)	.425 (.035)	.268 (.088)
β_2^*	.263 (.111)	.243 (.112)	.236 (.035)	.236 (.034)	.237 (.034)	.263 (.039)
β_3^*	.258 (.136)	NA	NA	NA	NA	.255 (.135)

Point estimates (PE) and standard deviations (SD) from logistic regression with reduced and maximal models, meta-analysis and GENMETA estimation with $\beta_1^* = \beta_2^* = \beta_3^* = \log(1.3) \approx .262$. NA means there is no corresponding estimator.

parameters (θ_1 and θ_2) across the two studies. We assume the sample sizes for the two studies to be $n_1 = 500$ and $n_2 = 5000$, and that for the reference dataset to be $n = 300$.

From the results reported in Table 2.2, we observe that in this simulation setting the reduced models produce biased estimate for β_1^*, but not for β_2^*. The result is intuitive given that the omitted covariate X_3 is primarily correlated with X_1. As a result, standard meta-analysis was nearly unbiased for β_2^*, but not for β_1^*. Parameter estimates from the maximal model from study-I are unbiased for all parameters, but have much larger standard error compared to meta-analysis for estimation of β_2^*. The GENMETA estimator produced unbiased estimates for all parameters and at the same time has comparable efficiency as standard meta-analysis for estimation of β_2^*. These results highlight the desirable feature of the GENMETA estimator that it can effectively combine information across studies to minimize bias due to omitted covariates and yet utilize all the information available across the partially informative studies.

2.3.2 Heterogeneous Population

In this section, we consider simulation studies where the underlying assumption of the homogeneity of covariate distribution across populations may be violated in multiple different ways. As a bench mark for comparison, we will describe setting (I) as the same setting as the the one we simulate under homogeneous population. In the setting (II), we allow the means or/and variances to vary across the populations underlying the studies and reference sample, keeping the correlations to remain constant. Specifically, we assume the mean-vector for the three covariates can take one of three possible values: $\mu_h = (1, 1, 1)$, $\mu_m = (0.5, 0.5, 0.5)$ and $\mu_b = (0, 0, 0)$. Similarly, the variance-vector is also allowed to vary across three possible set of values: $\sigma_h^2 = (2, 2, 2)$, $\sigma_l^2 = (0.5, 0.5, 0.5)$, $\sigma_b^2 = (1, 1, 1)$. In the setting (III), we then allow the correlations among the covariates to vary across populations. Here we also allow three possible set of correlation vector ρ as $\rho_l = (0.2, 0.4, 0.0)$, $\rho_h = (0.4, 0.8, 0.2)$ and $\rho_b = 0.3, 0.6, 0.1)$. Finally, we consider simulation setting (IV), where we allow for potential different inclusion criteria across studies leading to possible violations of the assumption of homogeneity of the covariate distribution. Specifically, we first simulate an underlying study base using the setup described in simulation setup (I), and then for study-I we only keep individuals with $X_1 > -0.5$ and $X_2 < 0.5$, and in study-II we keep individuals with $X_1 > 0$. Finally, we consider an alternative simulation scenario where we assume the covariates are log-normally distributed by defining $X = \exp(W)$, where W is generated from multivariate normal distribution following the same settings as I-IV described above

When covariates were normally distributed, we observe that (see Table 2.3) the proposed method is not very sensitive to underlying assumption of homogeneity of covariate distribution. In the setting (II), where the mean or/and variances of the covariates are varied across the population, but correlations are kept fixed, there is virtually no bias. In setting (III), where correlations are varied, we observe more noticeable, but still small, biases in parameter estimates. In setting (IV), when the inclusion criteria are varied across studies, there is also very minimal bias. When covariates are log-normally distributed, however, we observe that (see Table 1 in Supplementary Material) the method could be more sensitive to the violation of the underlying homogeneity assumption. In particular, when the inclusion criteria varied across studies (setting IV), large bias in point estimate and low coverage probability are observed for estimation of coefficient associated with X_2, the covariate which is used to define fairly non-overlapping inclusion criterion across two studies. Notably, even in this scenario, minimal bias is observed for estimation of the other covariates in the model.

2.3.3 Power Evaluation of the Diagnostic Test ($T_{GENMETA}$)

We assess the power of the proposed test statistic, $T_{GENMETA}$ in the presence of heterogeneity in the regression parameters (β) across the studies. In the context of standard multivariate meta-analysis, where it is assumed that all the studies ascertain the same set of covariates, test for heterogeneity is performed using

Table 2.3: Robustness of GENMETA Estimation (Normally Distributed Covariates)

Setting	Study-I	Study-II	Study-III	Reference	β_i^*	Bias	SD (ESD)	RMSE	CR	AL
I	μ_b	μ_b	μ_b	μ_b	β_1^*	.001	.111 (.112)	.111	.947	.437
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	-.002	.098 (.099)	.098	.956	.389
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	.005	.096 (.098)	.096	.954	.382
II	μ_b	μ_h	μ_m	μ_b	β_1^*	.010	.103 (.104)	.103	.952	.405
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	-.006	.083 (.083)	.083	.954	.324
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	.005	.085 (.088)	.085	.956	.343
	μ_b	μ_b	μ_b	μ_b	β_1^*	.003	.139 (.136)	.139	.939	.529
	σ_b^2	σ_h^2	σ_l^2	σ_b^2	β_2^*	-.003	.084 (.086)	.084	.956	.335
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	.003	.112 (.111)	.112	.949	.431
	μ_b	μ_h	μ_m	μ_b	β_1^*	.013	.124 (.126)	.125	.946	.493
	σ_b^2	σ_h^2	σ_l^2	σ_b^2	β_2^*	-.006	.073 (.075)	.073	.958	.291
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	.005	.097 (.100)	.097	.949	.391
III	μ_b	μ_b	μ_b	μ_b	β_1^*	-.092	.142 (.151)	.169	.958	.579
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.019	.105 (.109)	.107	.963	.423
	ρ_b	ρ_b	ρ_b	ρ_h	β_3^*	.053	.120 (.129)	.131	.971	.495
	μ_b	μ_b	μ_b	μ_b	β_1^*	.035	.099 (.099)	.106	.917	.385
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.002	.096 (.096)	.096	.954	.377
	ρ_b	ρ_b	ρ_b	ρ_l	β_3^*	.012	.087 (.087)	.088	.944	.343
	μ_b	μ_b	μ_b	μ_b	β_1^*	.060	.113 (.113)	.128	.916	.443
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	-.001	.096 (.097)	.096	.955	.379
	ρ_l	ρ_b	ρ_h	ρ_l	β_3^*	-.006	.103 (.102)	.104	.944	.398
	μ_b	μ_b	μ_b	μ_b	β_1^*	.039	.130 (.132)	.135	.939	.515
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	-.006	.097 (.100)	.097	.958	.392
	ρ_l	ρ_b	ρ_h	ρ_b	β_3^*	-.027	.116 (.118)	.119	.944	.461
	μ_b	μ_b	μ_b	μ_b	β_1^*	-.036	.165 (.173)	.169	.957	.671
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.013	.103 (.109)	.104	.962	.424
	ρ_l	ρ_b	ρ_h	ρ_h	β_3^*	.003	.143 (.153)	.143	.959	.591
IV	$X_1 > -0.5,$ $X_2 < 0.5$	$X_2 > 0$	μ_b	μ_b	β_1^*	.014	.123 (.127)	.124	.961	.494
			σ_b^2	σ_b^2	β_2^*	-.008	.105 (.109)	.105	.965	.428
			ρ_b	ρ_b	β_3^*	-.001	.094 (.093)	.093	.958	.366

Biases, standard deviation (SD), estimated standard deviation (ESD), square roots of mean square errors (RMSE), coverage rates (CR), and average lengths (AL) of 95% confidence intervals of the GENMETA estimates using the study covariance estimators in the setting of logistic regression. In setting (I), data are simulated in ideal setting there the covariate distribution, characterized by mean, sd and correlation of normal variates, are assumed to same across all populations. In setting (II)-(IV), the assumption is violated by creating variations in mean/sd, correlations and selection criterion across the studies and reference sample. with different study and reference sample. The vector of covariate means, variances and correlations are denoted by denoted by $\mu_* = (\mu_1, \mu_2, \mu_3)$, $\sigma_*^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ and $\rho_* = (\rho_{12}, \rho_{23}, \rho_{13})$ for $* \in \{b, l, m, h\}$, where $\mu_b = (0, 0, 0)$, $\mu_m = (0.5, 0.5, 0.5)$, $\mu_h = (1, 1, 1)$; $\sigma_b^2 = (1, 1, 1)$, $\sigma_l^2 = (0.5, 0.5, 0.5)$, $\sigma_h^2 = (2, 2, 2)$ and $\rho_b = (0.3, 0.6, 0.1)$, $\rho_h = (0.4, 0.8, 0.2)$, $\rho_l = (0.2, 0.4, 0)$. Estimated standard deviation are obtained by the asymptotic formula (2.2) and used to construct 95% confidence interval.

standard multivariate Cochran's test-statistic in the form

$$Q = \sum_{k=1}^{K} (\hat{\beta}_k - \hat{\beta}_{meta})^T S_k^{-1} (\hat{\beta}_k - \hat{\beta}_{meta})$$

where $\hat{\beta}_{meta}$ is the usual multivariate meta-analysis estimate and S_k is the standard error of $\hat{\beta}_k$ for $k = 1, \ldots, K$. We will utilize Q as a benchmark to evaluate the power of $T_{GENMETA}$.

In all simulations, as before, we assume the existence of three separate studies and relationship between a binary outcome variable Y and three covariates (X_1, X_2, X_3) in each study follows the same logistic regression model of the form (2.7). However, instead of assuming a fixed set of β across all studies, we simulate different values of β from a normal distribution with mean $(\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*) = (\log 1.3, \log 1.3, \log 1.3)$ and variance $\sigma^2 I$, where the parameter $\sigma^2 > 0$ is varied to control the degree of heterogeneity across studies. As before, we assume that (X_1, X_2, X_3) follows a multivariate normal distribution with mean zero, unit variances and underlying correlations $\rho = \rho_{12} = 0.3, \rho_{13} = 0.6, \rho_{23} = 0.1)$ across all the three studies. We simulate data for the different studies from the above random-effects logistic regression model and then fit reduced models of the form (2.8) to the three different studies. In particular, we assume X_1 and X_2 included in Study-I, X_2 and X_3 in Study-II and X_1 and X_3 in Study-III. We fix the sample size of the studies at $n_1 = 3000$, $n_2 = 5000$ and $n_3 = 10000$ and vary sample size of the reference dataset. The level of the test is set to 5%. For the purpose of comparison, we also fit the maximal model to each study involving all three covariates and apply the standard Q-statistics for testing heterogeneity.

Comparison of power of $T_{GENMETA}$ and Q statistics shows that, as expected,

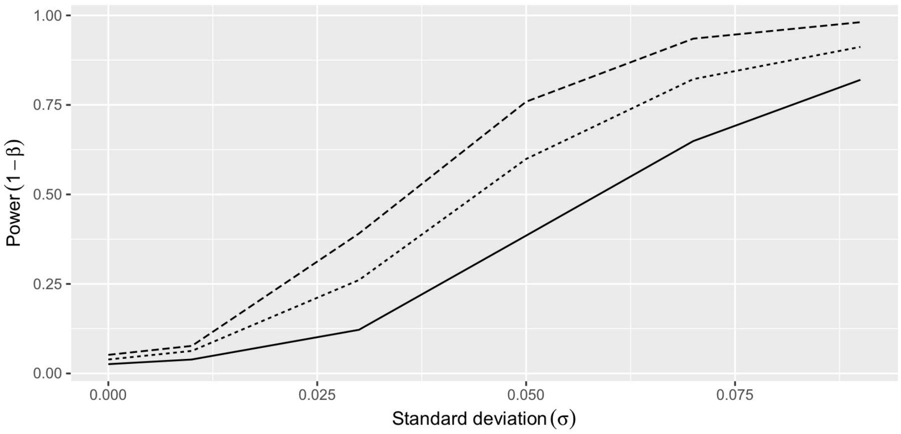


Figure 2.2: Power curves of simple multivariate meta-analysis test statistic (Q) and $T_{GENMETA}$ for simulated datasets. The long-dashed line is for the simple meta-analysis estimator. The solid and dotted lines are for GENMETA estimators with reference data sample sizes 100 and 500, respectively. Level of the test (α) is set to 0.05.

the power for both tests increases as a function of degree of heterogeneity, σ^2 (Figure 2.2). Clearly, $T_{GENMETA}$ suffers some loss of power as it handles missing covariates, but it retains substantial power, even with small reference dataset ($n = 100$), to remain practically useful.

2.4 Real Data Analysis

In this section, we illustrate an application of the proposed methodology to develop a model for predicting risk of breast cancer based on combination of different risk factors using data from multiple studies. The first study, the Breast Prostate Colorectal Cancer Cohort study (BPC3), includes a total of 7448 cases and 8812 controls, drawn from eight different underlying cohorts. Details of the study, including its recent application for the development of breast cancer risk prediction model, can be found elsewhere [114]. In the current analysis, we focus on the analysis of breast cancer risk associated with a selected set of factors, including family history, age at menarche, age at first birth and weight. The second study involves a dataset involving 1217 cases and 1616 controls from the Breast Cancer Detection and Demonstration Project (BCDDP). The study has been previously used to develop an updated version of the widely popular Breast Cancer Risk Assessment tool [33] to incorporate mammographic density, the areal proportion of breast tissue that is radiographically dense, known to be a strong risk factor for breast cancer. The dataset from the BCDDP study included mammographic density and number of previous breast biopsy, in addition to all the factors considered in the BPC3 data analysis. Let X denote the common set of covariates that are measured across both the studies and Z be the factors that are available only in BCDDP. The goal is to estimate parameters associated with an underlying logistic regression model that includes all of the different factors. While the BPC3 study is large in size and represents multiple populations, it has information on more limited number of risk factors. The BCDDP study, on the other hand, has information on extended set of risk factors, but is much smaller in size. A combined analysis of these two studies can potentially lead to more generalizable and precise estimate of risk parameters.

Throughout the analysis, we used a sample of 137 cases and 163 controls from the BCDDP study as the reference sample based on which the distribution of covariates are estimated. To maintain independence of the reference and study samples, we exclude the reference sample from the primary analysis of the BCDDP study that involved estimation of the log-odds-ratio parameters. Further, both the studies involve case-control sampling with similar case-control proportions. In general, if non-random sampling is used for selection of subjects in any of the studies, then the covariate distribution underlying the GENMETA estimating equation needs to be adjusted to account for the study design. In this application, because we had access to the the BCDDP study, we could adjust for the design effect by simply selecting a reference sample that includes cases and controls in similar ratio as the main studies. In general, however, the effect of non-random sampling design for the main studies may need to adjusted through careful weighting of subjects in the reference sample.

For each of the eight cohorts within the BPC3 study and for the BCCDP study, we first fit a reduced logistic regression model including X. All models included age as an additional cofactor and included study specific intercept parameters and age effects. Specifically, we consider underlying models in the form

$$(Y \mid X, \text{Age}, \text{study} = k) \sim \text{Bernoulli}((1 + \exp\{-(\theta_{0k} + \theta_{A_k} Age + \theta_X^T X)\})^{-1}). \tag{2.9}$$

We applied the diagnostic test for model violation to these datasets. We found the value of the test-statistic ($\hat{T}_{GENMETA}$) to be 59.01 and the corresponding p-value to be 0.366 under a $\chi^2_{(56)}$ distribution. Thus, it appears that the underlying model assumptions are unlikely to be grossly violated in this application.

First, to illustrate how the proposed GENMETA estimator compares to standard meta-analysis method, we consider estimating the common underlying parameters of interest θ_X using these two alternative methods. We fitted model (2.9) separately for each study and obtained estimates of the parameters and covariance matrices. Then, for the underlying common parameter of interests

θ_X, we conducted a standard multivariate meta-analysis using the corresponding subset of parameters estimates and covariance matrices. Alternatively, using the parameters estimates and variance-covariance matrices from the individual studies, and using the set aside BCDDP sample as the reference dataset to estimate the joint distribution of X and age, we estimated all of the parameters of model (2.9) using the GENMETA procedure. From the results reported in Table 2.4, we observe that in this setting, the meta-analysis and GENMETA estimators produce similar estimates as well as their standard errors across all the different risk-factors of interest. In one of the results stated earlier, we have seen theoretically that in an idealized setting where all the models and underlying populations are identical, the two estimators are asymptotically equivalent. It's encouraging to observe the close correspondence between the estimators in the data analysis, which includes a diverse set of studies that are likely to have significant heterogeneity across the underlying populations. In particular, for a number of the risk-factors (e.g family history), coefficient estimates were noticeably different across the two studies. When significant heterogeneity existed, the meta-analyzed estimates were pooled closer to those from the BPC3 study due to its large sample size.

Next, we turn our attention to the analysis of data from the BCDDP study using a maximal model that includes X and the additional covariates, mammographic density and number of previous breast biopsy. Comparison of the parameter estimates associated with X across the maximal and reduced model within the BCDDP study indicates major differences in the estimates of the coefficients associated with weight. In the maximal model, higher weight is found to be be much more strongly associated with increased risk of breast cancer.

The unmasking of the effect of weight in the maximal model is intuitive given that body weight and mammographic density is known to have strong negative correlation. Although not as dramatic, there are some differences in effects of age at menarchy and age at first birth between the maximal and reduced models, also possibly because of modest correlation of these factors with mammographic density and number of previous breast biopsy. The effect of family history, however, is almost identical across the two models.

Finally, we used the GENMETA method to combine estimates of the parameters of the maximal model from the BCDDP study and those from the reduced models from the eight BPC3 cohorts. We assumed an underlying maximal model of interest across the 9 studies in the form

$$(Y \mid X, Z, \text{Age}, \text{study=k}) \sim \text{Bernoulli}([1+\exp\{-(\theta_{0k}+\theta_{A_k}Age+\beta_X^T X+\beta_Z^T Z)\}]^{-1}).$$

We observe that GENMETA produces estimates of effect of family history and associated standard error very similar to those observed based on the standard meta-analysis of the reduced models across the nine cohorts. The estimate is pooled heavily towards the BPC3 study due to its large sample size. In contrast, the GENMETA estimates for weight are very similar to those observed from the maximal model only within the BCDDP study. These results are consistent with simulation studies, where GENMETA behaves similar to reduced model meta-analysis when omitted covariates do not cause notable bias. In contrast, when omitted covariates cause important bias, the GENMETA estimator is pooled towards estimates from maximal or more complete models that may be available from a restricted set of studies. The behavior of GENMETA for the two other covariates, age at menarchy and age at first birth, were in between, which is also intuitive given that we had observed their coefficients changed notably, but less dramatically, in the maximal model compared to the reduced model within the BCDDP study. The GENMETA parameter estimates and standard errors for the additional variables mammographic density and number of previous breast biopsy, were similar to those observed for the maximal model in the BCDDP, the only study which had information on these two factors. Thus, overall the data analysis illustrates that the GENMETA estimator behaves in a similar manner as meta-analysis for combining information across multiple possibly heterogeneous studies, but it has the added flexibility to effectively combine information from disparate models.

Table 2.4: Combined analysis of BCDDP and BPC3 study to develop a multivariate logistic regression model for breast cancer risk. For each cohort within BPC3 and for BCDDP, standard logistic regression model is applied for fitting reduced models including FH (family history), AMEN (age at menarche), AFB (age at first live birth) and WT (weight). Parameter estimates of the reduced models across studies are then combined using standard meta-analysis (meta) or GMeta. For the BCDDP study, a maximal logistic model is fitted including additional covariates mammographic density (MD) and number of previous biopsy (NBIOPS). These estimates are then combined with with estimates of reduced model parameters from BPC3 studies to obtain GMeta estimates of the maximal model. Point estimates (PE) and standard errors (SE) are shown for each analysis. NA means there is no corresponding estimator. The variables analyzed include: FH: binary indicator of family history; AMEN1 and AMEN2: dummy variables associated with age-at-menarche categories ≥ 14, 12–13 and ≤ 11; AFB1 and AFB2: dummy variables associated with age-at-first-live-birth categories ≤ 20, 21–29 and ≥ 30; WT1 and WT2: dummy variables associated with weight categories ≤ 62.6, 62.6–73.1 and ≥ 73.1 in kilograms; NBIOPS: the number of biopsies coded as a conitunous variable and MD: the standardized mammographic density coded as a continuous variable.

	BCDDP		BPC3								Meta	GMeta	
	Maximal Model	Reduced Model	CPS2 Cohort	EPIC Cohort	MCCS Cohort	MEC Cohort	NHS Cohort	PLCO Cohort	WHI Cohort	WHS Cohort	Reduced Model	Reduced Model	Maximal Model
	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)	PE(SE)
FH1	.80(.14)	.80(.14)	.47(.13)	.29(.15)	.56(.19)	.41(.28)	.48(.08)	.39(.13)	.30(.06)	.28(.19)	.40(.04)	.42(.04)	.37(.08)
AMEN1	.11(.10)	.07(.10)	-.03(.14)	.02(.09)	-.19(.17)	-.09(.24)	.06(.09)	-.05(.12)	.13(.08)	.03(.17)	.04(.04)	.03(.04)	.04(.06)
AMEN2	.55(.15)	.45(.15)	-.09(.17)	.04(.12)	-.44(.23)	.35(.35)	.19(.10)	.03(.15)	.19(.09)	.14(.19)	.13(.05)	.13(.05)	.32(.08)
AFB1	.06(.14)	.18(.15)	.28(.17)	.12(.14)	-.08(.25)	.06(.17)	.39(.20)	.16(.14)	.19(.09)	.92(.23)	.21(.05)	.20(.05)	.05(.09)
AFB2	.29(.20)	.46(.20)	.73(.24)	.24(.17)	.35(.30)	.05(.26)	.36(.22)	.52(.22)	.44(.13)	.96(.28)	.38(.06)	.38(.07)	.21(.12)
WT1	.29(.11)	.09(.11)	.09(.14)	-.01(.09)	.22(.18)	.09(.17)	.21(.08)	.09(.13)	-.03(.08)	-.01(.14)	.08(.04)	.08(.04)	.31(.07)
WT2	.52(.13)	.10(.13)	.16(.14)	.24(.11)	.45(.19)	-.08(.18)	.10(.08)	.09(.13)	.18(.08)	-.16(.15)	.14(.04)	.14(.04)	.63(.09)
NBIOPS	.13(.09)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	.13(.10)
MD	.46(.05)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	.43(.06)

2.5 Discussion

The proposed method can be viewed as a natural extension of the traditional fixed effect meta-analysis method that is widely used in practice. Both simulation studies and data analysis demonstrate that the method not only provides theoretically valid and efficient inference in idealized conditions, but also can perform robustly in non-idealized settings. A critical element of the proposed method is the access to a reference dataset. While the ideal choice of the reference dataset will vary by applications, publicly available survey data, which collect information on a wide variety of factors, can be useful broadly. In fact, in large scale genetic association studies, use of reference samples, such as the 1000 Genome study, are commonly used for estimation correlation parameters across genetic markers in the genome [38, 39, 99]. For epidemiologic studies, good resources for reference dataset for the US population include the National Health Interview Survey [2, 11, 8] and the National Health and Nutrional Examination Survey [55, 72, 43, 79, 96], which routinely collect data on a wide variety of health and lifestyle related factors. If multiple studies coordinate through consortium effort, which is increasingly common in biomedical applications, then studies which have most complete information, at least on some sub-samples, can provide reference sample.

When information on all covariates are not available in a single reference sample, one may have to consider simulation for generating such data by combining information from multiple studies under some modeling assumptions. As the access to large reference dataset that is ideally representative of the underlying study populations can be difficult, we found two aspects of GENMETA to be appealing. First, the sample size for the reference dataset can be small relative to the study datasets and yet GENMETA can have reasonable efficiency. In fact, increasing the sample size for the reference dataset beyond certain threshold does not have an impact on the efficiency of GENMETA. Second, although technically the method requires all the populations underlying the studies and the reference dataset to be the same, in practice, the method can be robust to a reasonable degree of heterogeneity in distribution of covariates. However, it is possible to have a large bias when estimating coefficients associated with covariates that have been used to define widely varying inclusion criteria. When different studies follow very different designs it is best to obtain study-specific reference samples for estimating the underlying moment equations. Alternatively, it may be possible to modify a large reference sample by using study-specific sampling weights/inclusion criteria when estimating the moment equations. Dealing with study-specific covariates, such as centers within a study, can also pose challenges as information on such variables are not expected to be available from a common reference sample. We have illustrated in our data example that it is possible to deal with such variables by imposing additional independence assumptions from other factors. In general, such complications need to be dealt in a case-by-case basis and some study specific reference samples may be needed to avoid making strong assumptions. Further research is merited to explore these and other practical challenges in implementation of the proposed method.

In general, we believe caution is needed for interpretations and applications of models that may be developed by combining information from disparate models across multiple studies. A model developed from a single study with complete information, although may be inefficient and may lack generalizability, is more likely to be internally consistent and thus can provide valid etiologic inference even if it is not representative of the general population. On the other hand, etiologic interpretation of parameters can be difficult when the underlying model is developed using information across multiple studies that are potentially heterogeneous. For the development of predictive models, however, where the focus is not so much parameter interpretation, development of rich models by combining information across multiple studies and then validating such models in independent studies can be an appealing strategy. These and other practical issues related to model development using multiple data sources have been also discussed in several recent articles [160, 66, 35, 52].

We used generalized method of moments as the underlying inferential framework. Alternatively, inference could be also performed using empirical likelihood theory [135, 136, 28] exploiting the same set of moment equations as we propose. While in small sample, empirical likelihood estimators may perform better, implementation can be substantially more complex. Recently, a simulation based method has been also described for combining information on model parameters across disparate studies [138]. Computationally, the proposed method may also enjoy substantial advantages in dealing with complex models, such as those in high-dimensional settings, where repeated model fitting on simulated data is extensive. Further research is merited in multiple directions to increase the practical utility of GENMETA. It is possible that in some applications we may have information only on subsets of parameters underlying the fitted reduced models. It's an open question how such partial information can be used to set up the underlying moment equations in the GENMETA procedure. Ideally, to increase robustness of inference, the GENMETA procedure should use study specific reference sample for setting up the moment equations. For this purpose, it may be useful to develop strategies to combine information on a common reference sample with complete covariate information and data from individual studies that have partial covariate information.

Chapter 3

Analysis of Two-Phase Studies using Generalized Method of Moments

3.1 Introduction

Modern epidemiological studies often require collection of information on a large number of factors, including lifestyle and behavioral factors, social and environmental conditions, and biomarkers. Measuring certain factors, such as novel biomarkers or physical activity levels based on wearable devices, can be cost-prohibitive. The difficulty can be overcome by employing a two-phase sampling design where at phase-I, a relatively large number of individuals are sampled from a target population for the ascertainment of a set of inexpensive covariates. At phase-II, a small sub-sample is then judiciously selected, possibly stratified by disease status and covariate information collected at phase-I, for the ascertainment of more expensive covariates. Two-phase sampling was first introduced by Neyman in 1938 as an approach for stratification and gradually, it gained popularity in many other fields including epidemiology, econometrics and GWAS [125, 112, 132, 145, 154]. Several studies have illustrated the design and analysis of two-phase studies using the data from the National Wilms Tumor Study [42, 62, 14].

Existing methods for logistic regression analysis of two-phase epidemiological studies include weighted-likelihood [56] (WL) and conditional-likelihood [164, 75, 13] (CML), which essentially focus on the analysis of the phase-II data, after accounting for sampling probability through weights or offsets, respectively. Information from phase-I data in these methods can be incorporated through post-hoc estimation of sampling weights based on available covariates. A variety of methods have been proposed to analyze two-phase designs under a semi-parametric missing data framework, where no modeling assumption regarding distribution of covariates is required. Examples include methods based on estimated-likelihood [129, 26, 77], regression calibration [34], pseudo-score [30] , weighted likelihood with weights calibrated by various sample survey techniques [20] and semiparametric maximum likelihood [141, 19, 146, 98, 171, 137].

In this chapter, we address two major challenges associated with the existing methods. First, a variety of methods assume that the available phase-I data can be summarized into a finite number of strata and as a result, they cannot effectively utilize information available on continuous covariates at phase-I. For example, many researchers have proposed semiparametric maximum likelihood estimation, but these methods are only efficient under the assumption that the phase-I data can be summarized into finite strata [18]. Another challenge for the analysis of two-phase studies can arise in the setting of large consortium based studies that require data sharing. For example, large consortia have been formed for conducting GWAS of various diseases. In such consortia, studies often share individual-level data on samples (e.g. a case-control sample) which are genotyped, but individual-level data from the large underlying study (e.g. a cohort study) is not typically made available. In such a setting, it may still be possible to get some summary-level information from phase-I, such as estimates of parameters associated with a reduced model including some basic covariates. Thus, methods that can incorporate summary-level data from the phase-I data can facilitate the incorporation of two-phase design methodology in consortia setting.

We propose a method for the analysis of two-phase studies with a binary outcome where phase-I data can potentially involve numerous covariates, some of which could be continuous. We summarize the information from phase-I data through parameters associated with the fitting of a reduced logistic regression model. We then use the individual-level data from phase-II and estimates of the reduced model parameters from phase-I to set up a set of estimating equations for inference on parameters associated with an extended logistic regression model of interest. We use the generalized method of moment (GMM) techniques for parameter estimation and asymptotic inference. Through simulation study and real data analysis, we show that the proposed method has the same efficiency as SPMLE when the phase-I data are discrete and yet it provides more flexibility to efficiently incorporate richer phase-I data by controlling the complexity of the reduced model.

This chapter is organized as follows: in section 3.1.1, the notations and statistical formulation of the problem is described followed by asymptotic properties of the proposed estimator. In section 3.2, extensive simulations are conducted under different sampling designs to study the performance of the proposed method. In section 3.3, we illustrate applications of our method using data from the US National Wilms Tumor Study.

3.1.1 Model Formulation

Let us denote the outcome of interest by Y, a binary variable taking values 1 and 0, and the set of full covariates by X, where dimension of X is q_2. We assume the true relationship between Y and X is given by a full model , or sometimes referred to as an extended model, of the form,

$$P(Y = i|X = x) = \frac{exp(i\beta^T X)}{1 + exp(\beta^T X)}. \tag{3.1}$$

Our goal is to estimate and draw inference about β_0, the true value of β. Before we move onto the estimation procedure, we introduce the two-phase sampling design considered here.

3.1.2 Sampling Design

We assume at phase-I, N samples are randomly drawn from an underlying population on each of which Y and Z, a set of covariates of dimension q_1, are observed. We assume Z to be a subset of X, but it could also include surrogates of some components of X where Z does not have any effect on outcome Y, given X.

More specifically, we will assume $Pr(Y|X,Z) = Pr(Y|X)$. Let $S := S(Z)$ denote a set of stratifying variables in phase-I. From each of the strata defined by Y and S at phase-I, a random sub-sample is drawn in phase-II based on known selection probabilities denoted by $\pi(Y,S)$.

3.1.3 Method

We first propose to summarize the phase-I data through a reduced model and use the reduced model parameters to establish an estimating equation for the full model parameters. We fit a reduced model of the form,

$$Pr(Y = i|Z) = \frac{exp(\theta^T Z)}{1 + exp(\theta^T Z)}; i = 0, 1 \tag{3.2}$$

to the phase-I data.

We will denote $\hat{\theta}$ to be the maximum-likelihood estimator of θ and denote θ_0 as the asymptotic limit of $\hat{\theta}$. Then, irrespective of whether the reduced model (3.2) is correctly specified or not, we can write $E\{\mathcal{S}(Y,Z;\theta)\}|_{\theta=\theta_0} = 0$, where $\mathcal{S}(Y,Z;\theta)$ is the score function and the expectation is taken under the true data generating distribution. Assuming the maximal model (3.1) is correct and using the law of iterated expectation, we can rewrite the score equation as $E\{f(X,Z;\beta,\theta)\}|_{\beta=\beta_0,\theta=\theta_0} = 0$, where $f(X,Z,\beta,\theta) := \{expit(\beta^T X) - expit(\theta^T Z)\}Z$ [28]. While evaluating this equation, we estimate the distribution of X empirically from the individual-level phase-II data with inverse probability weighting to account for non-random sampling design. Hence, an asymptotically unbiased estimating function for β, based on summary-level data ($\hat{\theta}$) available from phase-I, is given by,

$$U_{1N}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i f(X_i, Z_i, \beta, \hat{\theta})}{\pi(Y_i, S_i)},$$

where, R_i is an indicator variable determining the selection of ith subject in phase-II. For rigorous derivation of the above estimating equation, see supplemental material.

Further, we propose to use the following estimating equation to incorporate data from phase-II:

$$U_{2N}(\beta) = \frac{1}{N}\frac{N}{n}\sum_{i=1}^{N} R_i[Y_i - expit\{\gamma(\beta)^T X_i\}]X_i$$

The above estimating equation corresponds to standard logistic regression score equation where the effect of non-random sampling is accounted through incorporation of offset parameter in the logistic model parameter as: $\gamma(\beta) = \beta + (\log\frac{\pi(1,s)}{\pi(0,s)}, 0^T)^T$ [15].

We define β_{GMM}, the GMM estimator in two-phase design, to be the minimiser of the quadratic form, $Q_N(\beta) = U_N^T(\beta)\hat{C}U_N(\beta)$, where, $U_N(\beta) = (U_{1N}^T(\beta), U_{2N}^T(\beta))^T$ and $\hat{C}$ is a positive semi-definite matrix. Mathematically, $\hat{\beta}_{GMM} := \text{argmin}_\beta Q_N(\beta)$. From now on, for simplicity, we denote $\hat{\beta}_{GMM}$ by $\hat{\beta}$.

Let the limiting value of $\frac{n}{N}$ be λ, where we assume $\lambda \in (0,1)$. Let $\Psi(Y, X, R; \beta_0, \theta_0) = (\Psi_1^T, \Psi_2^T)^T$ denote the influence function of $U_N(\hat{\boldsymbol{\beta}})$, where $\Psi_1 := \Psi_1(Y, X, R; \beta_0, \theta_0) = \frac{Rf(X,\beta_0,\theta_0)}{\pi(Y,S)} + \{Y - expit(\theta_0^T Z)Z\}$, $\Psi_2 := \Psi_2(Y, X, R; \beta_0, \lambda) = \lambda^{-1}R\mathcal{S}(Y, X; \beta_0)$. Following the well established theory of GMM [71, 51, 81], we have the following theorem:

Theorem 3.1.1 (Consistency and Asymptotic Normality of $\hat{\beta}$). *Suppose the positive semi-definite matrix* $\hat{C} \xrightarrow{P} C$*. Then, under the regularity conditions (RC1-RC4) provided in appendix,* $\hat{\beta} \xrightarrow{P} \beta_0$*. Further, we have*

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, (\Gamma^T C\Gamma)^{-1}\Gamma^T C\Delta\Omega\Delta^T C\Gamma(\Gamma^T C\Gamma)^{-1})$$

where $\Omega = E(\Psi\Psi^T)$*,* $\Delta = \begin{pmatrix} I_{q_1} & 0 & I_{q_1} \\ 0 & \lambda^{-1} I_{q_2} & 0 \end{pmatrix}$ *and* $\Gamma = E\frac{\partial}{\partial\beta}U(\beta, \theta)|_{\beta=\beta_0,\theta=\theta_0}$*.*

The above asymptotic variance is minimized at the optimal C given by $C_{opt} = (\Delta\Omega\Delta^T)^{-1}$. Then, the optimal asymptotic variance is given by $(\Gamma^T(\Delta\Omega\Delta^T)^{-1}\Gamma)^{-1}$. We compute $\hat{\beta}$ using the following standard iterated GMM algorithm [70].

Algorithm:

(i) First we choose C to be an identity matrix and then minimize the quadratic form to get an initial estimate, $\hat{\beta}^{(1)}$.

(ii) Using the estimate obtained in step (i), we compute $\hat{C} = \hat{C}_{opt} = \{\hat{\Delta}\hat{\Omega}(\hat{\beta}^{(1)})\hat{\Delta}^T\}^{-1}$. With this $\hat{C}$, we minimize the quadratic form to obtain $\hat{\beta}^{(2)}$.

(iii) Iterate step (ii) with the estimate obtained in step (ii) till convergence.

For a rigorous proof of the theorem, see supplemental material.

3.2 Simulation Studies Resembling US National Wilms Tumor Data

We conduct simulations to gain insight into the results from real data analysis involving the NWTS study (see section 3.3). The data contains 4028 children diagnosed with Wilms Tumor, the most common form of kidney cancer in the pediatric age group, recruited in the third and fourth clinical trial of the National Wilms Tumor study. Details of the study can be found elsewhere [42, 62, 14]. The outcome variable of interest in this is study is relapse, a binary variable with 1 indicating that the patient's condition has deteriorated. The covariates of interest are: institutional histology (0 if favourable/1 if unfavourable); central histology (0 if favourable/1 if unfavourable); stage (0 if stage-I/1 if stage-II, 2 if stage-III and 3 if stage-IV) and age. There were two types of histology measurements available in the study. First, the institutional histology, i.e., the classification of the tumor into favorable and unfavorable, according to the pathologist at the hospital where the children were admitted for their treatment. Because the data came from many different hospitals, it's expected that the institutional histology is likely to be more error prone due to variations associated with subjective judgements from the different pathologists. Thus, the NWTG re-evaluated histology using a central pathologist recruited for the entire study which is referred to as central histology, the second measurement for histology available in the study.

Imitating this structure of the real data, we assume existence of four covariates, X_1, X_2, X_3, and X_4, where X_1 and X_2 are binary variables taking values 0 and 1; X_3 is an ordinal variable taking values 0,1,2 and 3; and X_4 is a continuous variable assumed to follow standard normal distribution. The covariates are simulated in a way such that the correlations among them are $(\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}) = (.73, .13, -.01, .09, .01, .27)$ and marginal probabilities for the discrete variables are: $Pr(X_1 = 1) = .9$, $Pr(X_2 = 1) = .89$ and $(Pr(X_3 = 0), Pr(X_3 = 1), Pr(X_3 = 2)) = (.39, .26, .23)$. These values are matched to those observed in real data. The algorithm underlying the simulation mechanism is described elsewhere [3]. Let $D = (D_1, D_2, D_3)$ denote the set of dummy variables constructed for coding the variable, X_3, in categorical form in the underlying models. We assume the relationship between Y and the covariates in the source population can be described by a logistic regression model of the form

$$Pr(Y = 1|X_2, D, X_4) = h(\beta_0+\beta_1 X_2+\beta_2^T D+\beta_3 X_4+\beta_4^T D\otimes X_2+\beta_5 X_2 X_4+\beta_6^T D\otimes X_4)$$

where, $\beta_1 = 1.16$, $\beta_2 = (\beta_{2A}, \beta_{2B}, \beta_{2C}) = (.60, .46, .81)$, $\beta_3 = .22$, $\beta_4 = (\beta_{4A}, \beta_{4B}, \beta_{4C}) = (.44, 1.03, 1.63)$, $\beta_5 = -.67$, $\beta_6 = (\beta_{6A}, \beta_{6B}, \beta_{6C}) = (.20, .33, .06)$ and $h(.) = (1 + (exp(.))^{-1})^{-1}$. These values are chosen by fitting the above model to the real data. The intercept parameter, β_0, is chosen to be -3.6 yielding a disease prevalence of 6%.

According to the above simulation scheme, we generated $10,000$ individuals in phase-I. We considered two sampling designs, a simple case-control design and a balanced design, for generating the phase-II sample. Under the case-control design, an equal number of samples are randomly drawn from the two strata, $Y = 1$ and $Y = 0$, respectively. In the balanced design, we draw random samples jointly stratified by Y and X_1 so that the resulting sample is balanced across both the levels of Y and the levels of X_1. Previous studies [13, 14] have shown that the balanced sampling design can gain efficiency over standard case-control sampling for estimation of parameters associated with a covariate for which balancing is achieved. During analysis of each data, we pretend that X_2 is observed only for those individuals who are selected at phase-II.

For data analysis using the proposed method, we considered the following two logistic regression models for summarizing the phase-I data.

$$\text{M1}: \; Pr(Y=1|X_1,D,X_4) = h(\gamma_0 + \gamma_1 X_1 + \gamma_D^T D + \gamma_4 X_4),$$

$$\text{M2}: \; Pr(Y=1|X_1,D,X_4) = h(\gamma_0+\gamma_1 X_1+\gamma_D^T D+\gamma_4 X_4+\gamma_5^T X_1 \otimes D+\gamma_6 X_1 X_4+\gamma_7^T D \otimes X_4),$$

where $\gamma_D = (\gamma_{D1}, \gamma_{D2}, \gamma_{D3})$, $\gamma_5 = (\gamma_{5A}, \gamma_{5B}, \gamma_{5C})$ and $\gamma_7 = (\gamma_{7A}, \gamma_{7B}, \gamma_{7C})$.

For the purpose of comparison, we also implemented a semiparametric maximum-likelihood estimator (SPMLE) where the phase-I data were summarized into discrete strata as stratum probabilities. The strata were defined by combination of (Y, X_1), the variables used for stratified sampling, and by X_3, which we included as a post-stratification variable to incorporate information on phase-I. The SPMLE was computed using the missreg package in R [148].

From the results shown in Table 3.1, we observe that GMM produce nearly unbiased estimates of the parameters, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ and their standard errors; and was able to maintain the coverage probabilities at the nominal level. We further observe that when the phase-I data were summarized using a more saturated model (M2), there was a very substantial gain in efficiency for the GMM estimator compared to SPMLE for covariates associated with X_4 and its interaction with other covariates. This highlights the desirable attribute of the GMM estimator that it can efficiently borrow information available from phase-I covariates. However, we also observed that when the phase-I data were summarized using a less saturated model (M1), the GMM estimator can lose substantial efficiency compared to the SPMLE for parameters associated with several covariates.

3.3 Application to US National Wilms Tumor Data

In this section, we demonstrate an application of our methodology to simulated two-phase data constructed from the real National Wilms Tumor study as described in section 3.2. Analogous to the study conducted earlier by Breslow and Chatterjee (1999) using this dataset, here we repeatedly simulated phase-II samples while keeping the phase-I sample to be fixed as the entire NWTS cohort [14].

Let D and S denote the outcome variable of interest, relapse status, and a stratum indicator variable for institutional histology, respectively. Let Z denote central histology and $W = (W_1, W_2, W_3)$ denote the set of dummy variables for stage, where $W = 0$ denotes stage-I. We assume the probability of relapse given all the covariates can be specified as,

$$Pr(Y = 1|S, Z, W, Age) = h(\beta_0+\beta_1 Z+\beta_2^T W+\beta_3 Age+\beta_4^T W\otimes Z+\beta_5 Z*Age+\beta_6 W\otimes Age) \tag{3.3}$$

, where we implicitly assumed that institutional histology has no information on relapse status given central histology and other covariates. Since we have all the variables measured in the full cohort, we assumed the ground truth to be the parameters associated with the model (3) fitted to the entire NWTS data. We simulated two-phase studies where we pretended that the institutional histology is available only at phase-II and we evaluated mean-squared errors of the GMM and SPMLE estimators around the ground truth.

For simulation of phase II data, we first classified all the subjects in phase-I into disjoint strata based on D and S, where the strata specific counts are provided in Table 3.2. We considered two different designs, case-control and balanced (see Table 3.2). Here, the balanced design is defined in a similar way as described by Breslow & Chatterjee [14] by sampling all the relapsed cases and all the patients with unfavorable histology. We simulated 1000 phase-II samples based on each of the designs with the associated sampling probabilities given in Table 3.2.

We summarized the phase-I data by fitting the following logistic regression model,

$$Pr(Y=1|Z_e,W,Age) = h(\theta_0+\theta_1 Z_e+\theta_2^T W+\theta_3^* Age+\theta_4^T W\otimes Z_e+\theta_5 Z_e*Age+\theta_6 W\otimes Age),$$

to the phase-I data where Z_e denotes institutional histology which is an error prone version of central histology, Z. From the simulated individual-level phase-II data and the information on parameter estimates, $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4, \hat{\theta}_5, \theta_6)$, obtained from the fitted model at phase-I, we estimated the regression parameters associated with model (3.3) using our proposed methodology. Although, the variance-covariance matrix associated with the phase-I model parameters can be estimated from the phase-II sample, however, we estimated it from phase-I data as we have access to the entire dataset in this application.

To compare the performance of the GMM estimator with the SPMLE estimator, we implemented the latter using the missreg package in R [148]. In the estimation procedure, the stage variable is used for post-stratification to incorporate as much information as available from phase-I. When we attempted to post-stratify based on both categories of age and stage, due to the sparsity in sample size in some of the cells of the cross-classified table, the SPMLE often failed to converge. Thus, in the final analysis, we implemented SPMLE with only stage as a post-stratification variable and hence incorporate the information on it from phase-I, but we incorporated the age information into the analysis only from the subjects included at phase-II.

We calculated the mean square error of the regression coefficients around the assumed ground truth. From Figure 3.1 and 3.2, we see substantial smaller MSE for the effect of age and its interaction with other covariates in both the designs. Also, we observed that under case-control design, the GMM produced larger MSE compared to SPMLE for model terms that did not include age effect. However, under the more efficient balanced design, this efficiency loss is modest specially considering the gain in efficienct for age-related terms.

3.4 Discussion

In this article, we have proposed a novel method for the analysis of two-phase studies which can incorporate information from complex multivariate phase-I data through summary-level parameters associated with fitted reduced models. We showed through extensive simulation studies and real data analysis that

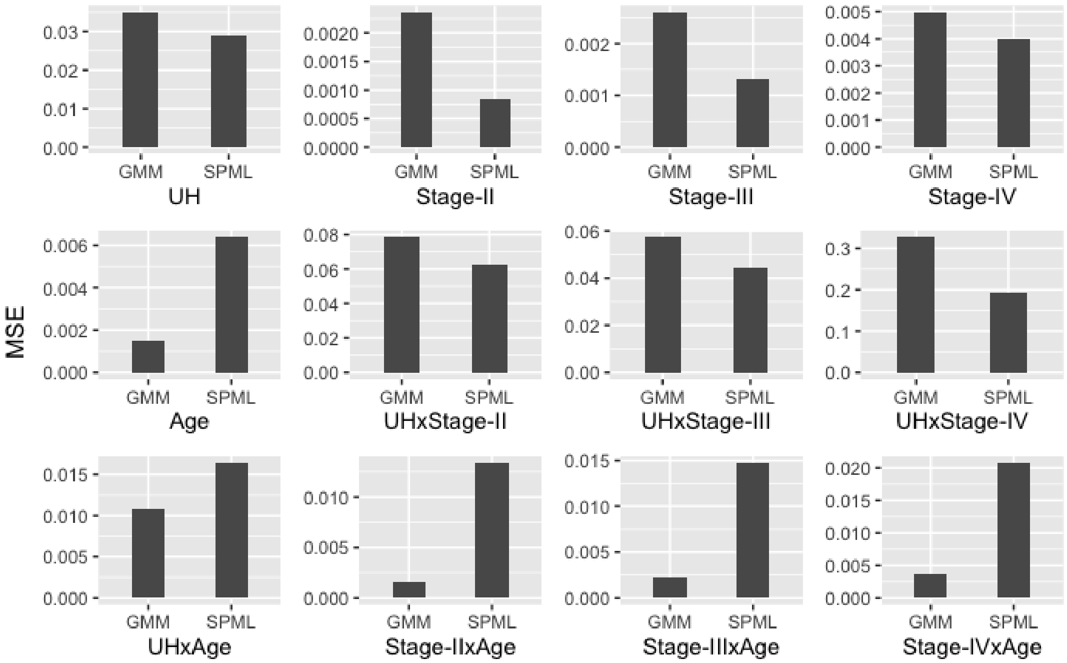


Figure 3.1: Mean square errors from real data analysis in case-control design

summarizing phase-I data through a set of parameters associated with an underlying reduced model, in contrast to summarizing the information into a set of strata, can lead to a more flexible and efficient way of utilizing the phase-I data in the analysis. The reduced model, however, should be made as saturated as possible as the size of the data permits. Use of a highly under-specified model can result in a substantial loss of efficiency (see Table 3.1).

We have considered scenarios where the selection probabilities were known by design. However, in large studies with complex designs, it may be considerably difficult to retrieve true selection probabilities. In such settings, one can estimate the selection probabilities in a post hoc fashion based on fitted parametric or semi-parametric models[59, 144, 21, 103, 131]. Further research

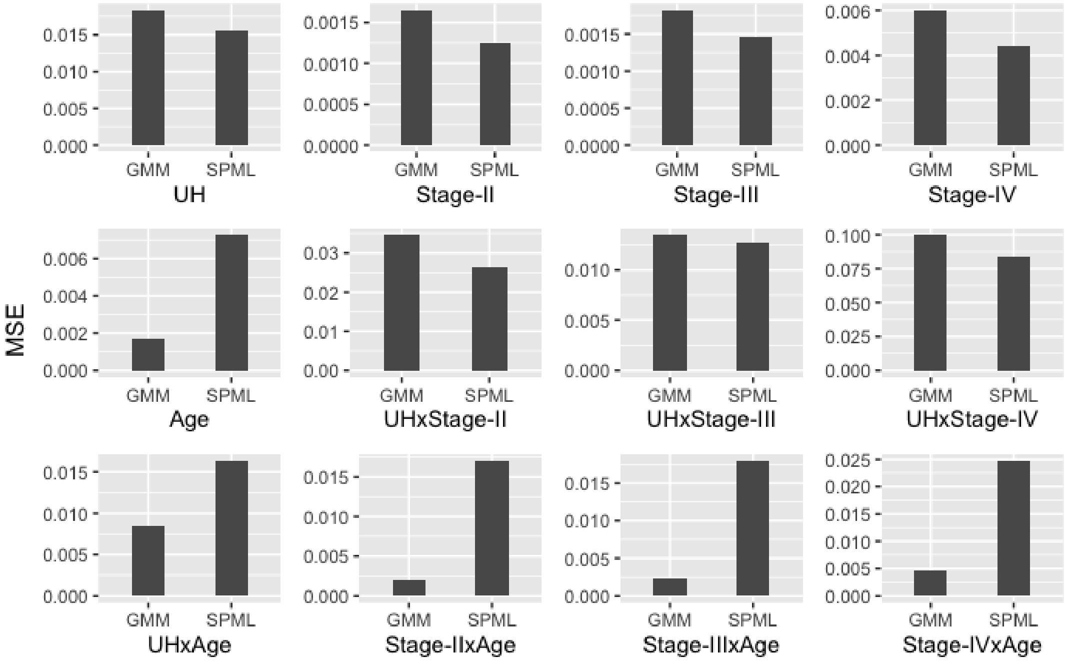


Figure 3.2: Mean square errors from real data analysis in balanced design

is merited to explore the impact of estimation of selection probabilities on efficiency of the proposed method.

Our method relies on the generalized method of moments framework for drawing an inference. Alternatively, inference would also be conducted using empirical likelihood (EL) theory [136, 135] using a similar set of estimating equations. Executing the EL approach may be notably complex in spite of enjoying small sample properties. Application of EL approach in two-stage outcome-dependent sampling designs have been discussed in recent articles [171, 137]. Computationally, the proposed method appreciates the benefits of an iterated re-weighted least squares algorithm.

We assumed the phase-I sample to be a random sample. However, there are many epidemiological studies that employ case-control sampling at phase-I itself

[15, 146, 19] or/and considers even more complex designs, such as multi-phase design [163] and partial questionnaire designs [158], all of which creates complex missing data by design. In those scenarios, one can amend the estimating equations accordingly to incorporate the particular design. Other extensions that merit future research include analysis of time-to-event outcomes based on hazard-based regression models under various two-phase sampling schemes for cohort studies, such as the case-cohort design [172, 104, 101, 159, 134].

Table 3.1: Simulation Results (Imitating Real Data Structure)

Design	Phase-I Covariates	Parameter	Bias(%)	SD (ESD)	CP	RE
Case-Control	(X_1, D, X_4)	β_1	.029	.33 (.31)	.95	.73
		β_{2A}	.003	.16 (.16)	.94	0.82
		β_{2B}	.004	.19 (.18)	.94	0.82
		β_{2C}	0.002	.24 (.23)	.95	.80
		β_3	0.003	.12 (.12)	.95	1.00
		β_{4A}	-0.015	.45 (.43)	.94	0.66
		β_{4B}	-.007	.46 (.43)	.94	0.68
		β_{4C}	.050	.57 (.53)	.95	0.64
		β_5	-.013	.17 (.16)	.95	1.09
		β_{6A}	-.009	.17 (.17)	.94	1.00
		β_{6B}	.003	.18 (.17)	.94	1.00
		β_{6C}	.003	.22 (.20)	.95	1.04
Balanced	(X_1, D, X_4)	β_1	-.03	.28 (.28)	.94	0.93
		β_{2A}	.004	.16 (.16)	.96	0.83
		β_{2B}	.007	.18 (.18)	.95	0.84
		β_{2C}	0.03	.22 (.22)	.94	.84
		β_3	1.29	.13 (.13)	.95	1.08
		β_{4A}	-.01	.37 (.36)	.95	0.86
		β_{4B}	-.01	.37 (.36)	.95	0.88
		β_{4C}	-.0008	.40 (.39)	.95	0.89
		β_5	-.02	.12 (.13)	.96	1.21
		β_{6A}	-.01	.18 (.18)	.95	1.04
		β_{6B}	-.001	.18 (.18)	.96	1.04
		β_{6C}	-.006	.21 (.20)	.94	1.05
Case-Control	$(X_1, D, X_4, X_1X_4, D \otimes X_2, D \otimes X_4)$	β_1	0.017	.298 (.284)	.94	0.89
		β_{2A}	0.005	.148 (.156)	.95	0.96
		β_{2B}	0.006	.175 (.173)	.94	0.93
		β_{2C}	0.022	.215 (.205)	.95	0.97
		β_3	0.001	.110 (.105)	.94	1.20
		β_{4A}	-.007	.388 (.375)	.93	0.88
		β_{4B}	-.003	.398 (.377)	.95	0.92
		β_{4C}	-.0008	.478 (.443)	.95	0.89
		β_5	-.002	.139 (.134)	.95	1.62
		β_{6A}	-.006	.145 (.139)	.95	1.40
		β_{6B}	.006	.149 (.141)	.95	1.45
		β_{6C}	.007	.176 (.162)	.94	1.65
Balanced	$(X_1, D, X_4, X_1X_4, D \otimes X_2, D \otimes X_4)$	β_1	-0.022	.274 (.275)	.94	0.97
		β_{2A}	0.011	.159 (.156)	.94	0.88
		β_{2B}	0.029	.176 (.178)	.96	0.87
		β_{2C}	0.052	.205 (.213)	.95	0.94
		β_3	-0.0006	.106 (.114)	.93	1.43
		β_{4A}	-0.019	.352 (.359)	.95	.91
		β_{4B}	-0.016	.349 (.355)	.96	0.93
		β_{4C}	-0.033	.378 (.389)	.94	0.96
		β_5	-0.001	.117 (.114)	.94	1.34
		β_{6A}	-0.011	.141 (.148)	.94	1.60
		β_{6B}	-0.008	.141 (.147)	.94	1.60
		β_{6C}	-0.006	.155 (.168)	.94	1.68

Biases, standard deviation (SD), estimated standard deviation (ESD), and coverage probabilities (CP) for GMM estimator. The last column shows relative efficiency (RE) with respect to SPMLE estimator.

Table 3.2: Wilms Tumor Data: Phase-I strata frequencies and Phase-II sampling design

Phase-I: Strata Frequencies			Phase-II Sampling Probabilities			
			Case-Control		Balanced	
Institutional Histology	Cases[a]	Controls	Cases[a]	Controls	Cases[a]	Controls
Favorable	415	3207	1	0.165	1	0.086
Unfavorable	156	250	1	0.165	1	1

[a] Cases are defined to be the relapsed ones.

Chapter 4

Genome-wide Interaction Scan Identifies Gene by Smoking interaction at 2q21.3 for Pancreatic Cancer Risk

4.1 Introduction

Pancreatic cancer is the seventh leading cause of cancer death worldwide [139]. In 2018, 458,918 new cases of pancreatic cancer were diagnosed, and 432,242 individuals died from this disease [139]. The incidence rates for pancreatic cancer has significantly increased since the mid-1990s in the United States and worldwide [61, 108]. Risk of pancreatic cancer increases dramatically with increasing age with the majority of cases diagnosed after 55 years of age [139]. Pancreatic ductal adenocarcinoma (PDAC) is the most common subtype and represents ≥

85% of total pancreatic cancer [139].

Inherited susceptibility plays an important role in pancreatic cancer risk as demonstrated by the high-risk of PDAC in individuals with a family history of pancreatic cancer, particularly those with multiple affected relatives [22]. Pathogenic variants in BRCA1, BRCA2, PALB2, ATM, CDKN2A, STK11, BRCA1 as well as DNA mismatch repair genes have been shown to increase risk of PDAC [76], with recent studies demonstrating that up to 10% of pancreatic cancer patient harbor pathogenic variants these genes [169]. Common variants also play an important role in PDAC, with array-based heritability estimates of up to 21.2% [32]. Our recent genome-wide association studies (GWAS) have identified over 18 regions with significant ($P - value \leq 5 \times 10^{-8}$) associated with PDAC. Associated gene regions include 1p36.33 (NOC2L), 2 independent loci at 1q32.1 (NR5A2), 2p13.3 (ETAA1), 3q29 (TP63), 3 loci at 5p15.33 (CLPTM1L- TERT), 7p12 (TNS3), 7p13 (SUGCT), 7q32.3 (LINC-PINT), 8q21.11 (HNF4G), 8q24.21(MYC), 9q34.2 (ABO), 13q12.2 (PDX1), 13q22.1 (KLF5), 16q23.1 (BCAR1), 17q12 (HNF1B), 17q25.1 (LINC00673), 18q21.32 (GRP) and 22q12.1 (ZNRF3) [4, 36, 93, 130, 166, 170].

In addition to inherited genetic factors, other risk factors for PDAC include cigarette smoking, diabetes, chronic pancreatitis, heavy alcohol use and excess body weight [139]. In particular, the association between smoking and PDAC is among the most well established with an estimated population attributable fraction in the United States of 12.1% [150]. Both case-control and cohort studies have demonstrated a close to 2-fold elevated risk among current smokers compared to never smokers [109, 10]. A pooled analysis of data from 12 case-control studies within the Pancreatic Cancer Case-Control Consortium (PanC4) showed that compared to never smokers, the odds ratio (OR) of PDAC was 1.17 (95% confidence interval [CI] 1.02-1.34) in former smokers and 2.20 (95% CI 1.71-2.83) for current cigarette smokers (15,16). A pooled nested case-control study within 12 cohorts in the Pancreatic Cancer Cohort Consortium (PanScan) showed an increased risk of PDAC among current smokers compared to never smokers (OR = 1.77, 95%CI: 1.38- 2.2) [109]. Although no overall association was observed among former smokers, former smokers who had quit less than 10 years had a significant elevated risk (OR=2.19, 95% CI 1.25-3.83) with the risk attenuating as cessation time increased and approached that of never smokers more than 15 years after quitting smoking [109].

Cigarette exposure has also been shown to cluster within families, and nicotine addiction has been shown to have a strong heritable component [167, 9]. To date, several genome-wide significant loci have been detected associated to distinct smoking related traits [78, 116, 105]. Studies include a recent GWAS analysis of over 1.2 million individuals that identified over 406 loci associated with the tobacco related traits [105]. Established associations include a cluster of nicotinic acetylcholine receptor (nAChR) genes highly expressed in the brain, CHRNA5-CHRNA3-CHRNB4 located on chromosome 15q24 [116, 105]. Despite the large number of associated loci, the common genetic variants altogether account only for 0.1% of the phenotypic variation in smoking cessation and 2.9% of the phenotypic variation in age at smoking initiation indicating highly polygenic nature of these traits [105].

Candidate gene studies, mostly related to carcinogen metabolism, DNA repair, oxidative stress and inflammation, have examined interactions by smoking for PDAC with inconsistent results [86]. A previous genome-wide gene-smoking interaction analysis for PDAC that included 2,028 cases and 2,109 controls from PanC4 did not show significant evidence of SNP by smoking interactions [153], however may have not had the power to detect modest effect sizes. A more comprehensive approach with a larger number of participants may detect associations not previously considered. Hence, in the present study we conducted genome-wide gene-by smoking interaction analysis of PDAC risk using genotype data from four prior GWAS studies conducted in the PanScan and PanC4 Consortia [4, 36, 93, 130, 166] and alternative statistical methods that have robust power for detecting gene-environment interactions [151].

4.2 Results

We conducted our study using 1000 Genomes imputed genotype data from the PanScan and PanC4 Consortia participants [4, 36, 93, 130, 166] with complete smoking status data (never, former, current). Our final analytic dataset was comprised of 6,769,447 common single nucleotide polymorphisms (SNP, MAF> 5%, INFO score >0.5) in 7,937 individuals with PDAC and 11,774 control individuals.

Figure 4.1 shows the Q-Q and Manhattan plot associated with genome-wide test for interaction using the Empirical Bayes (EB) method (see methods). Compared to the theoretical distributions, the lambda values for the interaction 0.93, showing reasonable control of type-I errors. Figure C.2 in Appendix C shows the Q-Q and Manhattan plots for the constrained maximum-likelihood (CML) and UML unconstrained maximum-likelihood (UML) methods.

We found a genome-wide significance interaction between smoking and SNPs

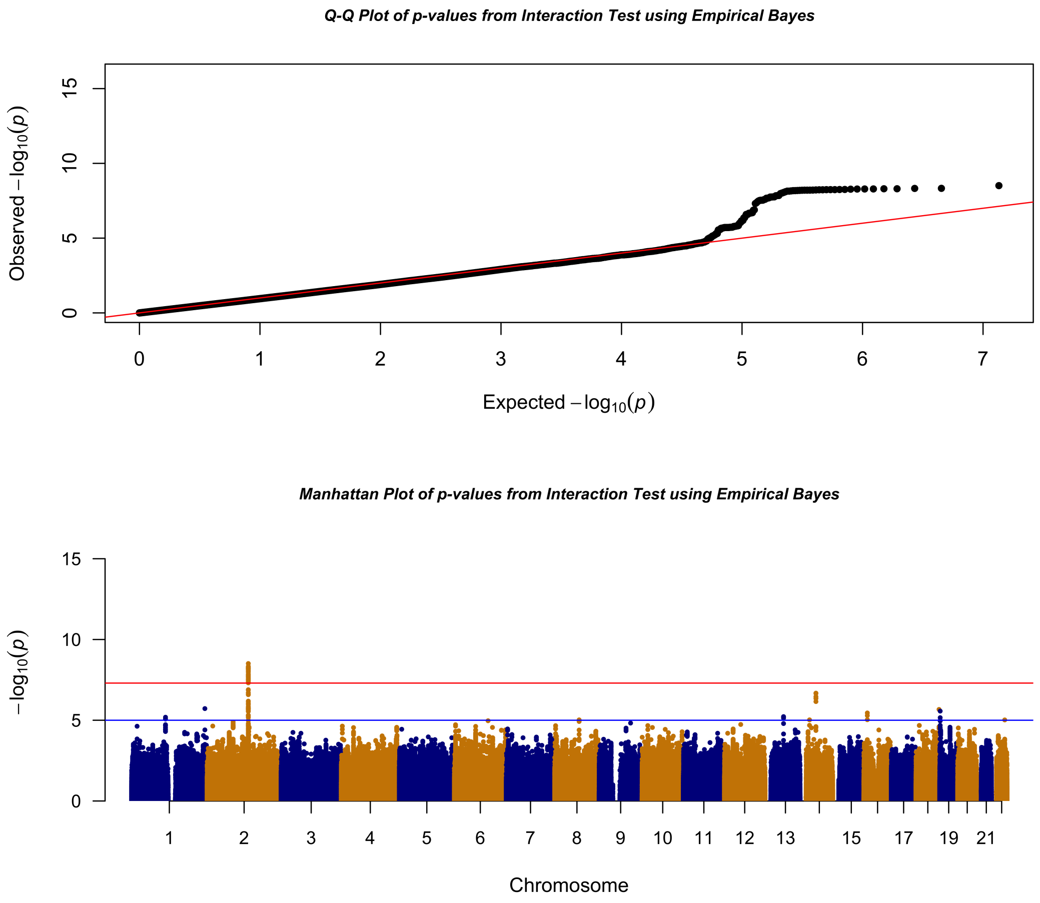


Figure 4.1: Q-Q and Manhattan Plots of Interaction Analysis using Empirical Bayes Approach

located in a region on chromosome 2q21.2 (Figure 4.2, [161]). The EB and CML methods detected a genome-wide significant interaction with p-values, 3.08×10^{-9} and 2.7×10^{-9}, respectively (Table 4.1). Evidence was also seen using the UML method though at below the genome-wide significance threshold. SNP rs1818613 provided the most significant evidence of interaction using both the EB and CML methods and more than 40 additional SNPs within the $\sim$ 100Kb region of high LD ($r^2 \geq 0.8$) also showed evidence of interaction (2

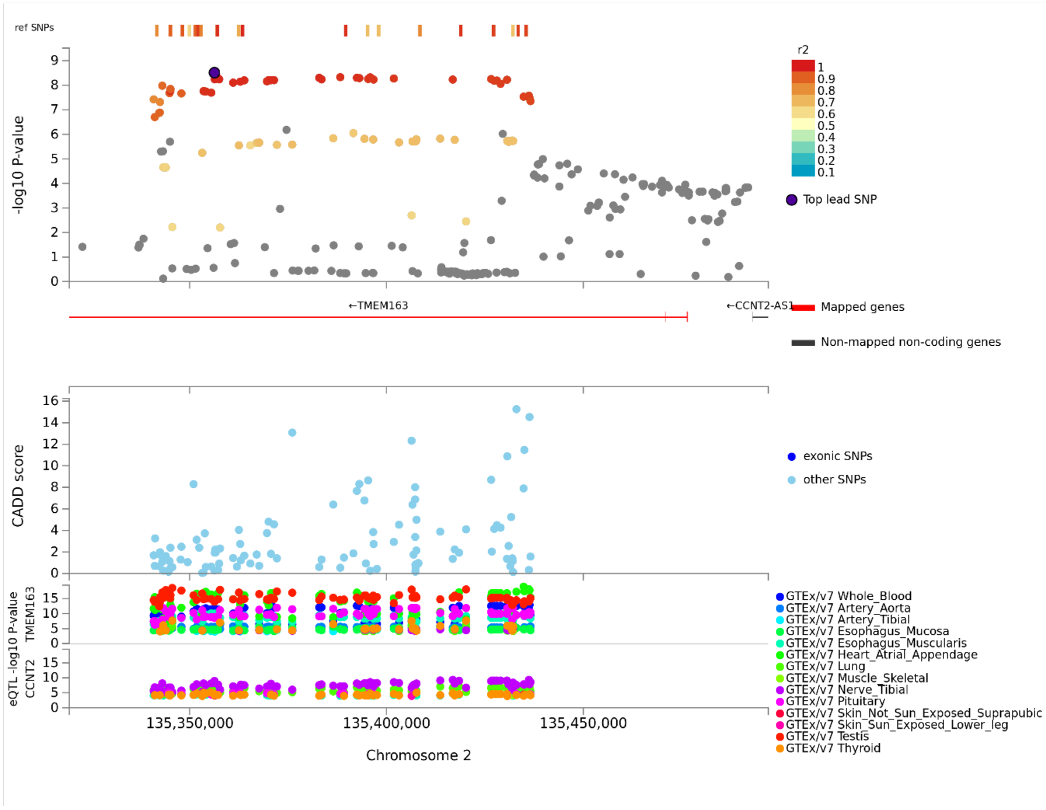


Figure 4.2: Locus plot of 2q21.3 region for the interaction GWAS of pancreatic cancer by smoking using FUMA [161]. a. Extended region of the $TMEM163$ locus that prioritizes genes $TMEM163$ and $CCNT2$. b. Zoomed in regional plot of $TMEM163$ locus with GWAS interaction P-values (SNPs are colored based on r^2), Combined Annotation Dependent Depletion (CADD score), and eQTL P-value. eQTLs are plotted per gene and colored based on tissue types.

degree-of-freedom, interaction EB p-value $\leq 5 \times 10^{-8}$) (Figure2). This region is located in intron 5 of transmembrane protein 163 gene ($TMEM163$) and 100kb upstream of transcription factor cylin T2 ($CCNT2$) (Figure 4.2). Compared to the G allele, the minor allele T of rs1818613 (Table 4.1) was associated with reduced risk of PDAC in never smokers (per allele EB-OR=0.87, 95% CI 0.82-0.93, $p-value = 0.001$), had a null effect in former smokers (OR=1.00, 95%CI 0.91-1.07, p-value= 0.94) and was associated with an increased risk of PDAC

Table 4.1: Region with genome wide significant evidence for SNP by smoking interaction on risk of PDAC

Chromosome Physical Position SNP Ref/Effect Alleles Ref Allele Frequency Imputation Quality Gene	Analytical Method	Odds Ratio for rs181613 (95% Confidence Interval) P-value			InteractionP-value#
		Never Smokers	Former Smokers	Current Smokers	
2q21.3 135356285 rs1818613 G/T 0.39 0.99 *TMEM*163(intronic)	CML	0.87 (0.82, 0.92) 1.19×10^{-6}	0.97 (0.91,1.02) 0.24	1.16 (1.07,1.25) 3.3×10^{-4}	2.7×10^{-9}
	EB	0.87 (0.82,0.93) 0.001	1.00 (0.93,1.07) 0.94	1.25 (1.12,1.40) 1×10^{-4}	3.08×10^{-9}
	UML	0.89 (0.84,0.96) 2.04×10^{-5}	0.99 (0.92,1.06) 0.74	1.17 (1.08,1.28) 2.6×10^{-4}	1.02×10^{-6}

Physical position in Build 37: CML, Constrained maximum-likelihood; EB, Empirical Bayes; UML, Unconstrained maximum-likelihood

# Based on 2 degrees of freedom chi-square test. Analysis was adjusted for age, sex, ancestry (via principle components) and for PanScan study phase and site)

among current smokers (OR 1.25, 95%CI 1.12-1.40, p-value= 1×10^{-4}) . Thus, there was evidence of qualitative interaction between rs1818631 and PDAC by cigarette smoking status. This pattern was fairly consistent across all of the three methods and GWAS studies (PanScan and PanC4) (Appendix C, Table C.1).

4.2.1 Established GWAS regions

We also examined whether smoking status modified the associations of the 18 independent previously identified GWAS SNPs for PDAC in European populations [4, 36, 93, 130, 166]. Overall, there was no interaction by smoking for any of these regions (all UM and EB p-values ≥ 0.05 and only one region had SNPs that has p-values borderline significant ($0.05 > \text{p-value} > 0.005$) under the CML method. (Appendix C, Table C.2).

Expression quantitative trait locus (eQTL) and co-localization analysis:

Using data from the Genotype-Tissue Expression (GTEx) Project, we examined if there was evidence for eQTLs with genes in this region and if any observed eQTLs colocalized with our SNP-by-smoking interactions results suggesting possible target gene(s) underlying this association. SNPs in this region, including rs1818613, were significantly associated with a differential expression of TMEM163 with the most significant showing carriers of the T allele increased expression in heart atrial appendage (P-value $= 1.6 \times 10^{-14}$), whole blood (p-value $= 3.2 \times 10^{-14}$), esophagus muscularis (p-value $= 1.0 \times 10^{-14}$) and pituitary (p-value$= 2.9 \times 10^{-9}$) and decreased expression in testis (p-value $=1.0 \times 10^{-14}$) tissue. In addition, there was significant evidence of decreased CCNT2 expression with the T allele in tibial nerve tissue (p-value$=1.1 \times 10^{-9}$) and lung tissue (p-value$=1.5 \times 10^{-7}$) (Figure 4.3).

We then conducted colocalization analysis to determine if there was support for a common SNP(s) underlying both of these highly significant associations. We used two methods, co-loc [58], which tests the hypothesis that a single causal SNP underlies both the eQTL and SNP by smoking association results and eCAVIAR [74], which allows for multiple shared causal signals. In both analyses, the posterior probability of a shared signal was extremely high. The most significant evidence was for rs842357, which is in strong LD with rs1818613 ($r^2 = 0.94$), and also had significant evidence of interaction with smoking (EB p-value $= 1.75 \times 10^{-08}$)(Figure 4.2). The A allele of rs842357 was associated with decreased expression of TMEM163 in heart atrial appendage, tibial nerve, and stomach compared with the T allele. The eCAVIAR posterior probability of a shared locus underlying both the SNP-by-smoking association and eQTL

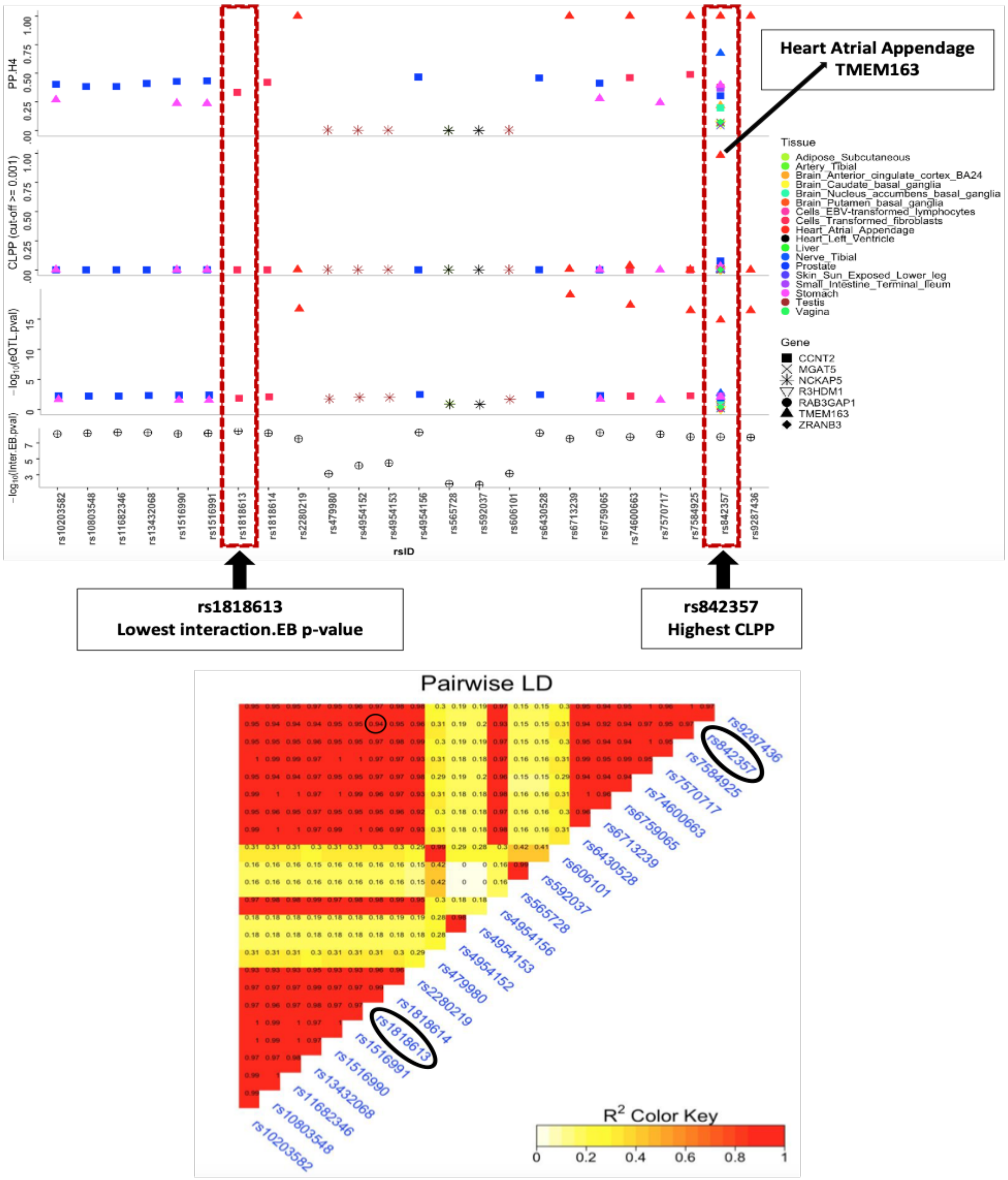


Figure 4.3: Results of colocalization analysis using eCAVIAR and co-loc. SNPs with colocalization probability (CLPP) ≥ 0.001 are shown in this plot. PP.H4 denotes posterior probability of having a common causal snp across eQTL and SNP X Smoking loci.

results for heart atrial appendage was 0.98. Additional evidence of colocalization (CLPP > 0.01) was observed for rs842357 for TMEM163 in brain anterior cingulate cortex BA24 and for CCNT2 in prostate, cells transformed fibroblasts, and small intestine terminal ileum (Figure C2). Interestingly, the A allele of rs843257 associated with decreased TMEM163 expression and increased CCNT2 expression.

4.3 Methods

4.3.1 Study sample

Study participants were selected from four previously conducted GWAS from the Pancreatic Cancer Cohort Consortium and the Pancreatic Case Control consortia. Details of these studies have been previously published [4, 36, 93, 130, 166]. Our study was based on 9,038 primary PDAC cases (ICD-O-3 code C250-C259) and 12,389 controls free of PDAC. Participants with non-exocrine pancreatic tumors were excluded (histology types 8150, 8151, 8153, 8155 and 8240). We only included participants of European ancestry to avoid confounding by population stratification. Each participating study obtained informed consent from participants and approval from their local Institutional Review Board. The Johns Hopkins School of Medicine and the National Cancer Institute's Special Studies Institutional Review Board approved the consortia study.

4.3.2 Genome-wide association genotyping data

Genotyping was conducted in four phases, PanScan I, PanScan II, PanScan III and PanC4. The PanScan studies were genotyped at the Cancer Genomics Research Laboratory (CGR) of the National Cancer Institute (NCI) of the National Institutes of Health (NIH) and genotyped on the Illumina HumanHap series arrays (Illumina HumanHap550 Infinium II [4], Human 610-Quad [130] for PanScan I-II, respectively, and the Illumina Omni series arrays (OmniExpress, Omni1M, Omni2.5 and Omni5M) for PanScan III [166]. PanC4 was genotyped on the Illumina HumanOmniExpressExome-8v1 array at the Johns Hopkins Center for Inherited Disease Research (CIDR) [36]. Details on imputation and quality controls prior to meta-analysis have been previously published [93]. In brief, for each study SNPs with call rates $\leq$ 98%, MAF ≤ 0.05 and Hardy-Weinberg equilibrium p value, measured in controls, was $< 1 \times 10^{-6}$ were excluded. SNPS were pre-phased using SHAPEIT2 software [44]. Genotype imputation was conduct using IMPUTE2 [113] with the 1000 genomes Phase 3 [1] as reference panel. Imputation was conducted separately for the PanScan I/II studies, Panscan III Study and PanC4 GWAS study. After imputation, we retained only SNPS with an imputation quality score > 0.5 and MAF was > 0.05. Data from the PanScan and PanC4 GWAS studies are available through dbGAP (accession numbers phs000206.v5.p3 and phs000648.v1.p1, respectively). Our final analysis included 6,769,447 variants.

4.3.3 Smoking and demographic assessment

Smoking status was assessed through self-report, proxy report or in-person interviews [109, 10]. We used the most recent accessed smoking status for the cohort studies [109]. For the case-control studies, smoking status at diagnosis (for cases) or when the questionnaire was administered (for controls)[10]. For these analyses smoking was categorized as never, former, and current smoker. Never smokers were individuals who smoked less than 100 cigarettes in their lifetime or less than 6 months. Former smokers were individuals who reported quitting cigarette smoking $>$ 1 year prior to the administration of the questionnaire. Current smokers were individuals who reported current smoking at the time of the questionnaire or who reported quitting cigarette smoking within the past year. Data on age, sex, and other possible confounders were collected from questionnaires at baseline from each cohort study and when smoking was assessed from the case-control studies [109, 10].

4.3.4 Statistical analyses

We used three alternative methods to evaluate the gene by smoking interaction, namely UML, CML, or EB [151]. The UML method corresponds to standard logistic regression analysis of case-control studies which allow the joint distribution of underlying covariates of the model to remain completely unspecified. The CML method, on the other hand, exploits as assumption of independence between SNP and smoking status in the underlying population [29]. The method, similar to the case-only method [133], can gain in efficiency for making inference in interaction parameter and yet it can be used to test or estimate all of the parameters of an underlying logistic model, including the main effect of a SNP and the exposure of interest. The EB method is an intermediate between the two methods above and allows data adaptive relaxation of the gene-environment independence assumption. Because the EB procedure provides a good compromise between bias and variance [121, 120], we use this as the primary method for evaluating the GWAS interaction while we used the other two methods for sensitivity analysis.

The association analysis was conducted using CGEN software (Version 3.5.0) (`https://dceg.cancer.gov/tools/analysis/cgen`), an R package for logistic regression analyses of SNP-environment interactions [68], using the 'snp.score' option in order to incorporate the genotype probabilities from the imputed data in the analysis. This option implements a score test (JScore), which tests for the join effect of gene and gene by environment interaction under a logistic regression model.

The analyses were first conducted separately for PanScan and PanC4 and results were combined using meta-analysis. Smoking was included as categorical dummy variable with never smokers as reference. Effect of each SNP was modeled under an additive model. Imputed SNPs were incorporated through expected dosage using snp.score function of the CGEN package [68]. Interaction between SNP and smoking was modeled using two parameters (current vs never) and the other for (former vs never). Each SNP genotype was coded using an underlying dosage model, coded in terms of observed/impute allele counts. The analysis was adjusted for age in decade, sex, and the top eigenvectors (5 for PanScan and 9 for PanC4) from principal components analysis (PCA) to control for ancestry. In addition, PanScan analyses were adjusted for study and geographic region of individual based upon parental study. For each SNP we obtained the one-step maximum-likelihood estimate of SNP and SNP-smoking interaction effects along with the associated variance-covariance matrix from the SNP-score function [151]. We implemented a fixed-effect meta-analysis using these summary statistics. Meta-analysis was performed separately for joint effect of SNP and SNP by smoking, and interaction effect only. Based on the meta-analyzed estimates, we performed a 2 degree-of-freedom tests for SNP by smoking interaction terms of the model as a way of identifying novel SNPs/regions the effect of which may be modified by smoking. In addition, we performed 3 degree-of-freedom joint tests (47) that simultaneously tests for both the main effect of a SNP and two SNP by E interaction terms. P-values less than 5×10^{-8} for the 2 degree-of-freedom tests were considered statistically significant.

Expression quantitative trait locus (eQTL) analysis: We examined eQTL to assess the cis effects of the rs1818163 genotype and corresponding 2q21region on gene expression across multiple tissues using the NIH Genotype-Tissue Expression (GTEx) v7 (https://gtexportal.org/home/) [65]. In addition, we created regional plots of the locus 2q21.3 region surrounding the rs1818163 genotype using the SNP2GENE function of FUMA [161]. FUMA incorporates information from multiple biologic resources and data repositories for functional annotation [161].

Colocalization analysis: For each SNP, we first meta-analyzed estimate of interaction parameters across current and former smokers to obtain a single estimate of SNP by smoking interaction under a dose-response model for smoking with never former and current coded as 0, 1 and 2, respectively. We used single statistics to summarize the evidence of interaction of individual SNPs with respect to smoking status in the colocalization analysis for the ease of interpretation of final results. Estimates of interaction separately by smoking categories indicate the dose-response model is adequate. To perform colocalization analysis, we matched the reference and alternate allele across our genome-wide interaction study and the eQTL results from GTEx v7. We performed colocalization analysis using two methods, co-loc and eCAVIAR [58, 74]. For eCAVIAR, we investigated the locus by considering 500Kb upstream and downstream of the most significant SNP, rs1818613, from the genome-wide interaction scan. In addition, we chose genes with at least one significant variant and set the maximum number causal variants to 3.

4.4 Discussion

We observed a qualitative interaction by cigarette smoking status for genetic variation and PDAC risk in a large LD block on chromosome 2 (2q21.3) in intron 5 of $TMEM163$ and upstream of $CCNT2$ such that alleles were associated with increased risk among current smokers and a decreased risk among never smokers. The pattern of the interaction was consistent across three analytical methods that rely on different assumptions regarding independence between the genetic variation and smoking exposure. The results were also consistent across the individual PDAC GWAS studies. Given the qualitative nature of this interaction, it is not surprising we did not observe an association in this region in our previous GWAS which did not stratify by smoking as the differing associations for smokers and non-smokers would result in no overall association. To the best of our knowledge, this is the largest gene-by-smoking interaction study for PDAC conducted to date.

The TMEM163 gene is conserved across many vertebrate species; it is highly expressed in specific brain regions and neuronal populations (glutamatergic and γ-aminobutyric acid (GABA)-ergic) [24], and is modestly expressed in other tissues including the pancreas, pituitary, and testis [65]. TMEM163 is a zinc binding and transporter protein involved in cellular zinc homeostasis and whose putative interaction with other zinc transporters and role in health and disease is not well understood [41]. Zinc mediates a wide range of cellular processes and alternations in its homeostasis can disrupt cellular function [5]. Dysregulation of other zinc transporters has been observed in PDAC such that zinc transporter upregulation has been associated with enhanced cancer cell migration and worse patient prognosis [5]. Interestingly, a genome-wide association study in the Finnish population reported an association in the same 2q21.3 region in an intron of $TMEM163$ gene for nicotine withdrawal in heavy smokers sampled from the population-based Finnish Twin Cohort study [78]. Of the three most significant SNPs described in this Finnish study, two (rs74865979 and rs62171406) were not present in 1000G. The third variant (rs75435861) was present in our sample with a MAF of 0.16 compared with 0.094 in Finnish population [78]. This SNP was in relatively low LD ($r^2 = 0.24, D' = 0.91$) with the associated SNPs in the present study and the evidence of interaction in PDAC susceptibility between this SNP and smoking was much weaker (EB P-value=0.0002) compared with that we observed for the lead SNPs. Germline variation in 2q21 region has also been associated with Parkinson's disease [122] and hematocrit concentrations [94] in populations of European ancestry, as well as, Type 2 diabetes in Asian Indians [152] and a Mongolian population in China

[6] .

Our colocalization results strongly support a single locus at the 2q21.3 region that underlies the qualitative interaction we observed for PDAC by cigarette smoking and the differential expression of TMEM163 and CCNT2 in several tissue types but not in pancreas tissue, which may imply the importance of gene-regulation beyond the pancreatic gland. The qualitative interaction suggests either a single mechanism that has inverse effects in never smokers compared with current smokers or given the role of this region in regulating multiple proteins, the protective effect observed in never-smokers is overwhelmed by a second risk-increasing mechanism in the context of cigarette smoking exposures.

Any hypothesis regarding the underly mechanism of the observed qualitative interaction between genetic variation at the 2q21.3 region and cigarette smoking is speculative. As mentioned above, the Finnish GWAS study linked variation inTMEM163 to nicotine dependence [78] and the increased PDAC risk in smokers may be due to differences in smoking behavior. In the Indian GWAS study, variation in on 2q21 at rs998451 within the TMEM163 region was associated with decreased plasma insulin concentrations and Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) (p-value < 0.008) with experimental studies supporting $TMEM163$ playing a possible role in zinc homeostasis in β-cells and insulin secretion, [152, 27]. Metals found in cigarette smoke, such a cadmium, have been shown to compete with other zinc transporters (e.g. metallothionein and ZIP8) and increase chronic toxicity [73]. It is possible a similar process could be contributing to the increased risk in smokers and qualitative interaction that we observe. Long term cigarette smoke spreads smoke related chemicals systemically in the bloodstream to target organs [126] and tobacco smoke inhalation causes pancreatic inflammation and damage to β-cells [126, 165]. Heavy smoking is a known risk factor for pancreatitis, diabetes, and PDAC [109, 10, 123, 64] and it is plausible that at least part of the interaction that we observe with the TMEM163 region may be related to pancreatogenic disease processes in smokers [126, 53] that are not present in never smokers.

In conclusion, we identified a qualitative interaction for PDAC by cigarette smoking status at 2q21.3 in intron 5 of the TMEM163 region. The co-localization results and eQTLs for $TMEM163$ and $CCNT2$ provides evidence of the importance of this gene region. Further studies are needed to replicate our observed association. In addition, studies are needed to understand functional mechanisms that could contribute to the qualitative interaction that we observe in smokers and never smokers and the clinical significance of our findings.

Chapter 5

Conclusions and Future Work

In this concluding chapter, we would like to summarize the contributions of this thesis and the possible impacts in public health, and delineate some of the potential future work.

5.0.1 Contribution

The key contributions are:

(i) **Methodology :** We developed a unified and general statistical framework for integrating disparate data sources. We developed an asymptotic theory of the proposed estimator based on the standard semiparametric theory behind the GMM approach, taking into account the uncertainty coming from the external studies. Although the framework is applied to breast cancer and kidney cancer, however, it can be used to other cancers, traits, or/and diseases where the outcome is binary or continuous.

(ii) **Software :** We developed an iteratively re-weighted least squares algorithm (IRWLS) and implemented it in the software for ease of use. It is well-known that IRWLS algorithm provides an easy way to approximately evaluate L^1 norm, which is considered to be more robust than L^2 norm. The code is accounted for scalability using standard vectorization tricks in R and standard sparse multiplication packages from Rcpp. We incorporated a function named GENMeta.plot, that graphically displays the estimates from the studies and the GENMeta estimate along with their 95% CI in the form of a dynamic series of forest plots for each covariate. The current version of the software is developed for linear and logistical regression models.

(iii) **Efficiency in two-phase studies :** Stratifying continuous variables in phase-I of a two-phase design can be ad-hoc and cumbersome. Classification of phase-I data that has a mixture of continuous and categorical variables can lead to some strata with limited number of individuals or even lead to empty strata. In such scenarios, the semi-parametric likelihood estimator can be computationally intractable. The proposed framework provides an alternate solution where all the phase-I data is captured in the form of parameters associated with a model and thus increases the efficiency of the parameters associated with continuous (or/and continuous-related) variables observed in phase-I.

(iii) **Applications in pancreatic cancer :** We carried out a genome-wide gene-by-environment(GxE) scan on the most extensive study on pancreatic cancer to date. Using existing GxE methods [151, 121, 29], we found novel SNP by smoking interactions at the genome-wide significance level in chromosome 2 associated with pancreatic cancer. This can help in identifying the underlying biological pathway(s) leading to pancreatic cancer.

Risk prediction is one of the vast areas in the scientific domain, which has critical applications in public health. The first essential step in risk prediction is to develop the risk prediction model. We believe the proposed methodology will aid in producing accurate and precise estimates of the parameters associated with the model. Eventually, this will help to stratify risk better to identify individuals with a higher risk for early detection of disease and help in making well-informed decisions for any intervention.

Inherited genetic changes and cigarette smoking are both known to play a significant role in the etiology of pancreatic cancer [93, 91, 92, 155]. Therefore, it is essential to study how the interaction between cigarette smoking and SNP is associated with pancreatic cancer risk. We found a susceptibility locus on 2q21 (long arm of chromosome 2) where pancreatic risk is modified by smoking status. Future studies are needed to explore this association in other smoking-related cancers and to understand the biological mechanism of such association. The study of discovering such associations (gene by smoking in our study) can provide insights into the underlying biological pathways leading to cancer (pancreatic cancer in our study) and thus better inform in making public health strategies for cancer prevention.

5.0.2 Future Work

There are potential extensions of the proposed methodology and many exciting applications of it with some modifications. We describe a few of them that are of particular interest to us:

(A) **Natural extensions :** We assumed fixed effect sizes across the studies. However, we would like to modify the framework under a random-effects model [85, 90]. MCMC type techniques can be used to estimate the posterior distribution that will be needed to evaluate the estimating equations [60]. Many researchers are often interested in modeling the time to an event instead of just looking at the binary outcome. It will be an exciting extension of our method to survival outcomes [40, 12]. GENMeta relies on the assumption that all the parameter estimates associated with the reduced models are provided to us or available from the literature. However, in some situations, some of the parameter estimates might not be reported in the literature and therefore, might be challenging to obtain. In those situations, it will be interesting to explore how to integrate such information in developing the maximal model.

(B) **Developing a model that includes high dimensional genetic factors :** In large consortia, GWAS are performed on various diseases. Due to a large number of cases, those case-control studies estimate the association of an SNP and a disease efficiently compared to cohort studies. Multiple web-based platforms provide easy access to these estimates and their standard errors. However, these associations might not be adjusted for other risk factors apart from primary demographic factors and genetic principal components. Extensive prospective cohort studies like UKBiobank collect information on a rich set of risk factors, including genetic factors. An interesting problem that we are currently working on is how to build a richer model on the UK Biobank population containing all the genetic and non-genetic risk factors of interest using GMM-LASSO approach by combining the information from large GWAS which provides a more precise estimate on SNP-disease association [57, 111, 106].

(C) **Multiple Outcomes :** The Breast Cancer Association Consortium (BCAC) is an international consortium encompassing 84 epidemiological and clinical breast cancer studies (http://www.b-cast.eu/eligibility/bcac/) [118, 110] . The study has rich information on genome-wide panel of SNP markers, epidemiologic risk factors, and a variety of tumor characteristics that can be potentially used to classify the disease into clinically distinct subtypes. A primary goal of the study is to understand how different genetic and epidemiologic risk factors are associated with the risk of breast cancer with specific subtypes of interest. However, various studies have varying levels of tumor characteristics information. Specifically, some studies have coarser information on breast cancer, such as only the indicator of the disease as yes/no or only have a subset of the relevant tumor characteristics. In contrast, some other studies have information on all relevant tumor characteristics, such as estrogen (ER) and progesterone (PR) hormone receptor status, human epidermal growth factor receptor status (HER2), stage, histology, and grade. This gives rise to different studies with varying levels of phenotypic information. For the study with the coarsest information on breast cancer, one usually fits a logistic regression model, whereas, for the study with finer level information, one usually fits a polytomous logistic regression. Suppose, we are provided with summary-level information (parameter estimates and their standard errors) from different fitted models corresponding to the different independent studies. Then, this setting fits into the GENMeta framework with appropriate key equations for synthesizing the summary level-information to generalize the results to a larger population.

(D) **Time-varying covariates :** The National Health and Nutrition Examination Survey (NHANES) data set(`https://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.html`) is a rich source of information on a variety of factors including demographic, socioeconomic, dietary and health-related factors in the US population of adults and children from multiple cohorts. It also has information on physical activity from the accelerometer device. For example, the accelerometry data from 2003- 2004 and 2005-2006 surveys containing minute-by-minute activity counts for seven days, which is more reliable than data from the questionnaire. This gives rise to a complex data structure with time-varying covariates. The National Center for Health Statistics(NCHS) has linked various surveys, including the NHANES, with death certificate records from the National Death Index(NDI), thus, providing information on mortality. One of the scientific questions of interest is to assess the association of risk factors with 5-year mortality in a larger population. Combining NHANES data over different years using GENMeta can be a potential application of GENMeta, especially given that in a more recent year, they have collected additional data, such as biomarker data, that were not available earlier. To be specific, suppose we are provided the parameter estimates and their standard errors corresponding to the risk factors(including activity counts) from the independent datasets in NHANES. We can now use the GENMeta methodology, modifying the key equations appropriately to incorporate the time-varying nature of the covariate, to estimate the parameter of the maximal model that is built by including all the covariates across the two studies.

References

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[2] P.F. Adams, G.E. Hendershot, and M.A. Marano. Current estimates from the national health interview survey, 1996. *Vital and health statistics*, 200:1–203, 1999.

[3] Anup Amatya, Hakan Demirtas, et al. Ordnor: An r package for concurrent generation of correlated ordinal and normal data. *Journal of Statistical Software*, 68(c02), 2015.

[4] Laufey Amundadottir, Peter Kraft, Rachael Z. Stolzenberg-Solomon, Charles S. Fuchs, Gloria M. Petersen, Alan A. Arslan, H. Bas Bueno-de Mesquita, Myron Gross, Kathy Helzlsouer, Eric J. Jacobs, Andrea LaCroix, Wei Zheng, Demetrius Albanes, William Bamlet, Christine D. Berg, Franco Berrino, Sheila Bingham, Julie E. Buring, Paige M. Bracci,

Federico Canzian, Françoise Clavel-Chapelon, Sandra Clipp, Michelle Cotterchio, Mariza de Andrade, Eric J. Duell, John W. Fox, Steven Gallinger, J. Michael Gaziano, Edward L. Giovannucci, Michael Goggins, Carlos A. González, Göran Hallmans, Susan E. Hankinson, Manal Hassan, Elizabeth A. Holly, David J. Hunter, Amy Hutchinson, Rebecca Jackson, Kevin B. Jacobs, Mazda Jenab, Rudolf Kaaks, Alison P. Klein, Charles Kooperberg, Robert C. Kurtz, Donghui Li, Shannon M. Lynch, Margaret Mandelson, Robert R. McWilliams, Julie B. Mendelsohn, Dominique S. Michaud, Sara H. Olson, Kim Overvad, Alpa V. Patel, Petra H.M. Peeters, Aleksandar Rajkovic, Elio Riboli, Harvey A. Risch, Xiao-Ou Shu, Gilles Thomas, Geoffrey S. Tobias, Dimitrios Trichopoulos, Stephen K. Van Den Eeden, Jarmo Virtamo, Jean Wactawski-Wende, Brian M. Wolpin, Herbert Yu, Kai Yu, Anne Zeleniuch-Jacquotte, Stephen J. Chanock, Patricia Hartge, and Robert N. Hoover. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature genetics*, 41(9):986–990, September 2009.

[5] Kyle J. Anderson, Robert T. Cormier, and Patricia M. Scott. Role of ion channels in gastrointestinal cancer. *World Journal of Gastroenterology*, 25(38):5732–5772, October 2019.

[6] Haihua Bai, Haiping Liu, Suyalatu Suyalatu, Xiaosen Guo, Shandan Chu, Ying Chen, Tianming Lan, Burenbatu Borjigin, Yuriy L. Orlov, Olga L. Posukh, Xiuqin Yang, Guilan Guilan, Ludmila P. Osipova, Qizhu Wu, and Narisu Narisu. Association Analysis of Genetic Variants with Type

2 Diabetes in a Mongolian Population in China. *Journal of Diabetes Research*, 2015:613236, 2015.

[7] J. B. Beckwith and N. F. Palmer. Histopathology and prognosis of wilms tumor results from the first national wilms' tumor study. 41(5):1937–1948.

[8] B. Bloom, R.A. Cohen, and G. Freeman. Summary health statistics for u.s. children: National health interview survey, 2009. *Vital and health statistics*, 247:1–82, 2010.

[9] Jason D. Boardman, Casey L. Blalock, and Fred C. Pampel. Trends in the Genetic Influences on Smoking. *Journal of health and social behavior*, 51(1):108–123, March 2010.

[10] C. Bosetti, E. Lucenteforte, D. T. Silverman, G. Petersen, P. M. Bracci, B. T. Ji, E. Negri, D. Li, H. A. Risch, S. H. Olson, S. Gallinger, A. B. Miller, H. B. Bueno-de Mesquita, R. Talamini, J. Polesel, P. Ghadirian, P. A. Baghurst, W. Zatonski, E. Fontham, W. R. Bamlet, E. A. Holly, P. Bertuccio, Y. T. Gao, M. Hassan, H. Yu, R. C. Kurtz, M. Cotterchio, J. Su, P. Maisonneuve, E. J. Duell, P. Boffetta, and C. La Vecchia. Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4). *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 23(7):1880–1888, July 2012.

[11] S. Botman and C.L. Moriarity. Design and estimation for the national health interview survey, 1995-2004. *Vital and health statistics*, 2:1–203, 2000.

[12] N. E. Breslow. Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975.

[13] N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, March 1988.

[14] N. E. Breslow and N. Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. 48(4):457–468.

[15] N.E. Breslow and K.C. Cain. Logistic regression for two-stage case control data. *Biometrika*, 75(1):11–20, 1988.

[16] N.E. Breslow and N. Chatterjee. Design and analysis of two phase studies with binary outcome applied to wilms tumor prognosis. *Appl. Statist.*, 48(3):457–468, 1999.

[17] N.E. Breslow and R. Holubkov. Maximum likelihood estimation for logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Statist. Soc. B*, 59(2):447–461, 1997.

[18] Norman Breslow, Brad McNeney, and Jon A. Wellner. Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics*, 31(4):1110–1139, August 2003.

[19] Norman E. Breslow and Richard Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. 59(2):447–461.

[20] Norman E. Breslow and Thomas Lumley. *Semiparametric models and two-phase samples: Applications to Cox regression.* Institute of Mathematical Statistics, 2013.

[21] Norman E. Breslow, Thomas Lumley, Christie M Ballantyne, Lloyd E. Chambless, and Michal Kulich. Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology. *Statistics in biosciences*, 1(1):32, May 2009.

[22] Kieran A. Brune, Bryan Lau, Emily Palmisano, Marcia Canto, Michael G. Goggins, Ralph H. Hruban, and Alison P. Klein. Importance of age of onset in pancreatic cancer kindreds. *Journal of the National Cancer Institute*, 102(2):119–126, January 2010.

[23] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[24] Jacqueline Burré, Herbert Zimmermann, and Walter Volknandt. Identification and characterization of SV31, a novel synaptic vesicle membrane protein and potential transporter. *Journal of Neurochemistry*, 103(1):276–287, October 2007.

[25] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and

Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, October 2018.

[26] R. J. Carroll and M. P. Wand. Semiparametric Estimation in Logistic Measurement Error Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):573–585, 1991.

[27] Shraddha Chakraborty, Shamsudheen Karuthedath Vellarikkal, Sridhar Sivasubbu, Soumya Sinha Roy, Nikhil Tandon, and Dwaipayan Bharadwaj. Role of Tmem163 in zinc-regulated insulin storage of MIN6 cells: Functional exploration of an Indian type 2 diabetes GWAS associated gene. *Biochemical and Biophysical Research Communications*, 522(4):1022–1029, February 2020.

[28] N. Chatterjee et al. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Am. Statist. Ass.*, 111(513):891–921, 2016.

[29] Nilanjan Chatterjee and Raymond J. Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92(2):399–418, June 2005.

[30] Nilanjan Chatterjee, Yi-Hau Chen, and Norman E. Breslow. A Pseudoscore Estimator for Regression Problems With Two-Phase Sampling. *Journal of the American Statistical Association*, 98(461):158–168, March 2003.

[31] Nilanjan Chatterjee, Zeynep Kalaylioglu, Roxana Moslehi, Ulrike Peters, and Sholom Wacholder. Powerful Multilocus Tests of Genetic Association

in the Presence of Gene-Gene and Gene-Environment Interactions. *The American Journal of Human Genetics*, 79(6):1002–1016, December 2006.

[32] Fei Chen, Erica J. Childs, Evelina Mocci, Paige Bracci, Steven Gallinger, Donghui Li, Rachel E. Neale, Sara H. Olson, Ghislaine Scelo, William R. Bamlet, Amanda L. Blackford, Michael Borges, Paul Brennan, Kari G. Chaffee, Priya Duggal, Manal J. Hassan, Elizabeth A. Holly, Rayjean J. Hung, Michael G. Goggins, Robert C. Kurtz, Ann L. Oberg, Irene Orlow, Herbert Yu, Gloria M. Petersen, Harvey A. Risch, and Alison P. Klein. Analysis of Heritability and Genetic Architecture of Pancreatic Cancer: A PanC4 Study. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 28(7):1238–1245, July 2019.

[33] J. Chen et al. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute*, 98(17):1215–1226, 2006.

[34] Yi-Hau Chen and Hung Chen. A Unified Approach to Regression Analysis under Double-Sampling Designs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(3):449–460, 2000.

[35] Wenting Cheng, Jeremy M. G. Taylor, Tian Gu, Scott A. Tomlins, and Bhramar Mukherjee. Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 0, 2018.

[36] Erica J. Childs, Evelina Mocci, Daniele Campa, Paige M. Bracci, Steven

Gallinger, Michael Goggins, Donghui Li, Rachel E. Neale, Sara H. Olson, Ghislaine Scelo, Laufey T. Amundadottir, William R. Bamlet, Maarten F. Bijlsma, Amanda Blackford, Michael Borges, Paul Brennan, Hermann Brenner, H. Bas Bueno-de Mesquita, Federico Canzian, Gabriele Capurso, Giulia M. Cavestro, Kari G. Chaffee, Stephen J. Chanock, Sean P. Cleary, Michelle Cotterchio, Lenka Foretova, Charles Fuchs, Niccola Funel, Maria Gazouli, Manal Hassan, Joseph M. Herman, Ivana Holcatova, Elizabeth A. Holly, Robert N. Hoover, Rayjean J. Hung, Vladimir Janout, Timothy J. Key, Juozas Kupcinskas, Robert C. Kurtz, Stefano Landi, Lingeng Lu, Ewa Malecka-Panas, Andrea Mambrini, Beatrice Mohelnikova-Duchonova, John P. Neoptolemos, Ann L. Oberg, Irene Orlow, Claudio Pasquali, Raffaele Pezzilli, Cosmeri Rizzato, Amethyst Saldia, Aldo Scarpa, Rachael Z. Stolzenberg-Solomon, Oliver Strobel, Francesca Tavano, Yogesh K. Vashist, Pavel Vodicka, Brian M. Wolpin, Herbert Yu, Gloria M. Petersen, Harvey A. Risch, and Alison P. Klein. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nature Genetics*, 47(8):911–916, August 2015.

[37] W. Chun, M.H. Chen, and E. Schifano. Statistical methods and computing for big data. *arXiv:1502.07989v2*, 2015.

[38] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[39] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[40] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[41] Math P. Cuajungco and Kirill Kiselyov. The Mucolipin-1 (TRPML1) Ion Channel, Transmembrane-163 (TMEM163) Protein, and Lysosomal Zinc Handling. *Frontiers in bioscience (Landmark edition)*, 22:1330–1343, March 2017.

[42] G. J. D'Angio, N. Breslow, J. B. Beckwith, A. Evans, H. Baum, A. deLorimier, D. Fernbach, E. Hrabovsky, B. Jones, and P. Kelalis. Treatment of wilms' tumor. results of the third national wilms' tumor study. 64(2):349–360.

[43] S.D. de Ferranti et al. Inflammation and changes in metabolic syndrome abnormalities in us adolescents: Findings from the 1988-1994 and 1999-2000 national health and nutrition examination surveys. *Clinical Chemistry*, 52(7):1325–30, 2006.

[44] Olivier Delaneau, Jonathan Marchini, 1000 Genomes Project Consortium, and 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, 5:3934, June 2014.

[45] R. Dersimonian and N. Laird. Meta-analysis in clinical-trials. *Control Clin Trials*, 7(3):177–88, 1986.

[46] R. Dersimonian and N. Laird. Meta-analysis in clinical trials revisited. *Contemp Clin Trials*, 45:139–145, 2015.

[47] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. 7(3):177–188.

[48] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials revisited. 45(0):139–145.

[49] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H. Granat, Tom White, Vincent T. van Hees, Michael I. Trenell, Christoper G. Owen, Stephen J. Preece, Rob Gillions, Simon Sheard, Tim Peakman, Soren Brage, and Nicholas J. Wareham. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*, 12(2):e0169649, February 2017.

[50] Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L. Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M. Smith. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726):210, October 2018.

[51] R. Engle and D. McFadden. *Handbook of Econometrics*. North Holland, 1994.

[52] Jason P. Estes, Bhramar Mukherjee, and Jeremy M. G. Taylor. Empirical bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences*, 2018.

[53] Nils Ewald and Reinhard G. Bretzel. Diabetes mellitus secondary to pancreatic diseases (Type 3c)–are we neglecting an important disease? *European Journal of Internal Medicine*, 24(3):203–206, April 2013.

[54] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014.

[55] J. Fang and M.H. Alderman. Serum uric acid and cardiovascular mortality the nhanes i epidemiologic follow-up study, 1971-1992. *JAMA Network*, 283(18):2404–10, 2000.

[56] W. Dana Flanders and Sander Greenland. Analytic methods for two-stage case-control studies and other stratified designs. 10(5):739–747.

[57] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[58] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5), May 2014.

[59] Peter B. Gilbert, Xuesong Yu, and Andrea Rotnitzky. Optimal Auxiliary-Covariate Based Two-Phase Sampling Design for Semiparametric Efficient Estimation of a Mean or Mean Difference, with Application to Clinical Trials. *Statistics in medicine*, 33(6):901–917, March 2014.

[60] W. R. Gilks, S. Richardson, David Spiegelhalter, S. Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, December 1995.

[61] Vanessa L. Gordon-Dseagu, Susan S. Devesa, Michael Goggins, and Rachael Stolzenberg-Solomon. Pancreatic cancer incidence trends: evidence from the Surveillance, Epidemiology and End Results (SEER) population-based data. *International Journal of Epidemiology*, 47(2):427–439, 2018.

[62] D. M. Green, N. E. Breslow, J. B. Beckwith, J. Z. Finklestein, P. E. Grundy, P. R. Thomas, T. Kim, S. J. Shochat, G. M. Haase, M. L. Ritchey, P. P. Kelalis, and G. J. D'Angio. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the National Wilms' Tumor Study Group. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 16(1):237–245, January 1998.

[63] Daniel M. Green, Norman E. Breslow, Giulio J. D'Angio, Marcio H. Malogolowkin, Michael L. Ritchey, Audrey E. Evans, J. Bruce Beckwith, Elizabeth J. Perlman, Robert C. Shamberger, Susan Peterson, Paul E. Grundy, Jeffrey S. Dome, Patrick R.M. Thomas, and John A. Kalapurakal. Outcome of patients with stage II/favorable histology wilms tumor with and without local tumor spill. a report from the national wilms tumor study group. 61(1):134–139.

[64] Julia B. Greer, Edwin Thrower, and Dhiraj Yadav. Epidemiologic and

Mechanistic Associations Between Smoking and Pancreatitis. *Current Treatment Options in Gastroenterology*, 13(3):332–346, September 2015.

[65] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, 348(6235):648–660, May 2015.

[66] Peisong Han and Jerald F. Lawless. Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica(online)*, 2017.

[67] Summer S. Han and Nilanjan Chatterjee. Review of Statistical Methods for Gene-Environment Interaction Analysis. *Current Epidemiology Reports*, 5(1):39–45, March 2018.

[68] Summer S. Han, Philip S. Rosenberg, Montse Garcia-Closas, Jonine D. Figueroa, Debra Silverman, Stephen J. Chanock, Nathaniel Rothman, and Nilanjan Chatterjee. Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting G-E independence for case-control data. *American Journal of Epidemiology*, 176(11):1060–1067, December 2012.

[69] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. 50(4):1029–1054.

[70] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

[71] L.P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[72] J. He et al. Risk factors for congestive heart failure in us men and women nhanes i epidemiologic follow-up study. *JAMA Network*, 161(7):996–1002, 2001.

[73] Seiichiro Himeno, Daigo Sumi, and Hitomi Fujishiro. Toxicometallomics of Cadmium, Manganese and Arsenic with Special Reference to the Roles of Metal Transporters. *Toxicological Research*, 35(4):311–317, October 2019.

[74] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V. Segrè, Xiao Li, Jong Wha J. Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, 99(6):1245–1260, December 2016.

[75] David A. Hsieh, Charles F. Manski, and Daniel McFadden. Estimation of Response Probabilities from Augmented Retrospective Observations. *Journal of the American Statistical Association*, 80(391):651–662, September 1985.

[76] Chunling Hu, Steven N. Hart, Eric C. Polley, Rohan Gnanaolivu, Hermela Shimelis, Kun Y. Lee, Jenna Lilyquist, Jie Na, Raymond Moore, Samuel O. Antwi, William R. Bamlet, Kari G. Chaffee, John DiCarlo, Zhong Wu, Raed Samara, Pashtoon M. Kasi, Robert R. McWilliams, Gloria M. Petersen, and Fergus J. Couch. Association Between Inherited

Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer. *JAMA*, 319(23):2401–2409, 2018.

[77] X. Joan Hu and Jerald F. Lawless. Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, 83(4):747–761, December 1996.

[78] Jenni Hällfors, Teemu Palviainen, Ida Surakka, Richa Gupta, Jadwiga Buchwald, Anu Raevuori, Samuli Ripatti, Tellervo Korhonen, Pekka Jousilahti, Pamela A. F. Madden, Jaakko Kaprio, and Anu Loukola. Genome-wide association study in Finnish twins highlights the connection between nicotine addiction and neurotrophin signaling pathway. *Addiction Biology*, 24(3):549–561, 2019.

[79] E.L. Idler and R.J. Angel. Self-rated health and mortality in the NHANES-I epidemiologic follow-up study. *American Journal of Public Health*, 80(4):446–452, 2011.

[80] Guido W. Imbens. One-step estimators for over-identified generalized method of moments models. pages 359–383.

[81] G.W. Imbens. Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20(4):493–506, 2002.

[82] John P. A. Ioannidis. Meta-analysis in public health: potentials and problems. 3(2).

[83] J.P.A. Ioannidis. Meta-analysis in public health: potentials and problems. *European Journal Public Health*, 15:60–61, 2005.

[84] D. Jackson, R. Riley, and I.R. White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[85] Jeroen P. Jansen, Bruce Crawford, Gert Bergman, and Wiro Stam. Bayesian Meta-Analysis of Multiple Treatment Comparisons: An Introduction to Mixed Treatment Comparisons. *Value in Health*, 11(5):956–964, September 2008.

[86] Rick J. Jansen, Xiang-Lin Tan, and Gloria M. Petersen. Gene-by-Environment Interactions in Pancreatic Cancer: Implications for Prevention. *The Yale Journal of Biology and Medicine*, 88(2):115–126, June 2015.

[87] M.I. Jordan. On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390, 2013.

[88] F.K. Kavvoura and J.P.A. Ioannidis. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Human Genetics*, 123(1):1–14, 2008.

[89] Robert W. Keener. *Theoretical Statistics: Topics for a Core Course.* Springer, 2010.

[90] NaNa Keum, Chung-Cheng Hsieh, and Nancy Cook. Random-effects meta-analysis of inconsistent effects. *Annals of Internal Medicine*, 161(5):379–380, September 2014.

[91] Alison P. Klein. Genetic Susceptibility to Pancreatic Cancer. *Molecular carcinogenesis*, 51(1):14–24, January 2012.

[92] Alison P. Klein, Sara Lindström, Julie B. Mendelsohn, Emily Steplowski, Alan A. Arslan, H. Bas Bueno-de Mesquita, Charles S. Fuchs, Steven Gallinger, Myron Gross, Kathy Helzlsouer, Elizabeth A. Holly, Eric J. Jacobs, Andrea Lacroix, Donghui Li, Margaret T. Mandelson, Sara H. Olson, Gloria M. Petersen, Harvey A. Risch, Rachael Z. Stolzenberg-Solomon, Wei Zheng, Laufey Amundadottir, Demetrius Albanes, Naomi E. Allen, William R. Bamlet, Marie-Christine Boutron-Ruault, Julie E. Buring, Paige M. Bracci, Federico Canzian, Sandra Clipp, Michelle Cotterchio, Eric J. Duell, Joanne Elena, J. Michael Gaziano, Edward L. Giovannucci, Michael Goggins, Göran Hallmans, Manal Hassan, Amy Hutchinson, David J. Hunter, Charles Kooperberg, Robert C. Kurtz, Simin Liu, Kim Overvad, Domenico Palli, Alpa V. Patel, Kari G. Rabe, Xiao-Ou Shu, Nadia Slimani, Geoffrey S. Tobias, Dimitrios Trichopoulos, Stephen K. Van Den Eeden, Paolo Vineis, Jarmo Virtamo, Jean Wactawski-Wende, Brian M. Wolpin, Herbert Yu, Kai Yu, Anne Zeleniuch-Jacquotte, Stephen J. Chanock, Robert N. Hoover, Patricia Hartge, and Peter Kraft. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. *PloS One*, 8(9):e72311, 2013.

[93] Alison P. Klein, Brian M. Wolpin, Harvey A. Risch, Rachael Z. Stolzenberg-Solomon, Evelina Mocci, Mingfeng Zhang, Federico Canzian, Erica J. Childs, Jason W. Hoskins, Ashley Jermusyk, Jun Zhong, Fei Chen, Demetrius Albanes, Gabriella Andreotti, Alan A. Arslan, Ana Babic, William R. Bamlet, Laura Beane-Freeman, Sonja I. Berndt, Amanda Blackford, Michael Borges, Ayelet Borgida, Paige M. Bracci,

Lauren Brais, Paul Brennan, Hermann Brenner, Bas Bueno-de Mesquita, Julie Buring, Daniele Campa, Gabriele Capurso, Giulia Martina Cavestro, Kari G. Chaffee, Charles C. Chung, Sean Cleary, Michelle Cotterchio, Frederike Dijk, Eric J. Duell, Lenka Foretova, Charles Fuchs, Niccola Funel, Steven Gallinger, J. Michael M. Gaziano, Maria Gazouli, Graham G. Giles, Edward Giovannucci, Michael Goggins, Gary E. Goodman, Phyllis J. Goodman, Thilo Hackert, Christopher Haiman, Patricia Hartge, Manal Hasan, Peter Hegyi, Kathy J. Helzlsouer, Joseph Herman, Ivana Holcatova, Elizabeth A. Holly, Robert Hoover, Rayjean J. Hung, Eric J. Jacobs, Krzysztof Jamroziak, Vladimir Janout, Rudolf Kaaks, Kay-Tee Khaw, Eric A. Klein, Manolis Kogevinas, Charles Kooperberg, Matthew H. Kulke, Juozas Kupcinskas, Robert J. Kurtz, Daniel Laheru, Stefano Landi, Rita T. Lawlor, I.-Min Lee, Loic LeMarchand, Lingeng Lu, Núria Malats, Andrea Mambrini, Satu Mannisto, Roger L. Milne, Beatrice Mohelníková-Duchoňová, Rachel E. Neale, John P. Neoptolemos, Ann L. Oberg, Sara H. Olson, Irene Orlow, Claudio Pasquali, Alpa V. Patel, Ulrike Peters, Raffaele Pezzilli, Miquel Porta, Francisco X. Real, Nathaniel Rothman, Ghislaine Scelo, Howard D. Sesso, Gianluca Severi, Xiao-Ou Shu, Debra Silverman, Jill P. Smith, Pavel Soucek, Malin Sund, Renata Talar-Wojnarowska, Francesca Tavano, Mark D. Thornquist, Geoffrey S. Tobias, Stephen K. Van Den Eeden, Yogesh Vashist, Kala Visvanathan, Pavel Vodicka, Jean Wactawski-Wende, Zhaoming Wang, Nicolas Wentzensen, Emily White, Herbert Yu, Kai Yu, Anne Zeleniuch-Jacquotte, Wei Zheng, Peter Kraft, Donghui Li, Stephen Chanock, Ofure Obazee, Gloria M. Petersen, and Laufey T. Amundadottir. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nature Communications*, 9(1):1–11, February 2018.

[94] Alexander M. Kulminski, Jian Huang, Yury Loika, Konstantin G. Arbeev, Olivia Bagley, Arseniy Yashkin, Matt Duan, and Irina Culminskaya. Strong impact of natural-selection-free heterogeneity in genetics of age-related phenotypes. *Aging*, 10(3):492–514, 2018.

[95] Prosenjit Kundu, Runlong Tang, and Nilanjan Chatterjee. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3):567–585, September 2019.

[96] J.S LaKind, M. Goodman, and D.Q. Naiman. Use of nhanes data to link chemical exposures to chronic diseases: A cautionary tale. *PLOS One*, 8(5):1295–1302, 2012.

[97] Vasileios Lapatas, Michalis Stefanidakis, Rafael C. Jimenez, Allegra Via, and Maria Victoria Schneider. Data integration in biological research: an overview. 22(1).

[98] J. F. Lawless, J. D. Kalbfleisch, and C. J. Wild. Semiparametric Methods for Response-Selective and Missing Data Problems in Regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):413–438, 1999.

[99] S.H. Lee, J. Yang, G.B. Chen, S. Ripke, E.A. Stahl, C.M. Hultman, P. Sklar, P.M. Visscher, P.F. Sullivan, M.E. Goddard, and N.R. Wray. Estimation of snp heritability from dense genotype data. *Am J Hum Genet.*, 93(6):1151–1155, 2013.

[100] Maurizio Lenzerini. *Data Integration: A Theoretical Perspective.*

[101] D. Y. Lin and Z. Ying. Cox Regression with Incomplete Covariate Measurements. *Journal of the American Statistical Association*, 88(424):1341–1349, 1993.

[102] D.Y. Lin and D. Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010.

[103] Roderick J. Little. Comment: Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, 22(2):171–174, May 2007. arXiv: 0710.5013.

[104] Dandan Liu, Tianxi Cai, Anna Lok, and Yingye Zheng. Nonparametric Maximum Likelihood Estimators of Time-Dependent Accuracy Measures for Survival Outcome Under Two-Stage Sampling Designs. *Journal of the American Statistical Association*, 113(522):882–892, 2018.

[105] Mengzhen Liu, Yu Jiang, Robbee Wedow, Yue Li, David M. Brazel, Fang Chen, Gargi Datta, Jose Davila-Velderrain, Daniel McGuire, Chao Tian, Xiaowei Zhan, Hélène Choquet, Anna R. Docherty, Jessica D. Faul, Johanna R. Foerster, Lars G. Fritsche, Maiken Elvestad Gabrielsen, Scott D. Gordon, Jeffrey Haessler, Jouke-Jan Hottenga, Hongyan Huang, Seon-Kyeong Jang, Philip R. Jansen, Yueh Ling, Reedik Mägi, Nana Matoba, George McMahon, Antonella Mulas, Valeria Orrù, Teemu Palviainen, Anita Pandit, Gunnar W. Reginsson, Anne Heidi Skogholt, Jennifer A. Smith, Amy E. Taylor, Constance Turman, Gonneke Willemsen,

Hannah Young, Kendra A. Young, Gregory J. M. Zajac, Wei Zhao, Wei Zhou, Gyda Bjornsdottir, Jason D. Boardman, Michael Boehnke, Dorret I. Boomsma, Chu Chen, Francesco Cucca, Gareth E. Davies, Charles B. Eaton, Marissa A. Ehringer, Tõnu Esko, Edoardo Fiorillo, Nathan A. Gillespie, Daniel F. Gudbjartsson, Toomas Haller, Kathleen Mullan Harris, Andrew C. Heath, John K. Hewitt, Ian B. Hickie, John E. Hokanson, Christian J. Hopfer, David J. Hunter, William G. Iacono, Eric O. Johnson, Yoichiro Kamatani, Sharon L. R. Kardia, Matthew C. Keller, Manolis Kellis, Charles Kooperberg, Peter Kraft, Kenneth S. Krauter, Markku Laakso, Penelope A. Lind, Anu Loukola, Sharon M. Lutz, Pamela A. F. Madden, Nicholas G. Martin, Matt McGue, Matthew B. McQueen, Sarah E. Medland, Andres Metspalu, Karen L. Mohlke, Jonas B. Nielsen, Yukinori Okada, Ulrike Peters, Tinca J. C. Polderman, Danielle Posthuma, Alexander P. Reiner, John P. Rice, Eric Rimm, Richard J. Rose, Valgerdur Runarsdottir, Michael C. Stallings, Alena Stančáková, Hreinn Stefansson, Khanh K. Thai, Hilary A. Tindle, Thorarinn Tyrfingsson, Tamara L. Wall, David R. Weir, Constance Weisner, John B. Whitfield, Bendik Slagsvold Winsvold, Jie Yin, Luisa Zuccolo, Laura J. Bierut, Kristian Hveem, James J. Lee, Marcus R. Munafò, Nancy L. Saccone, Cristen J. Willer, Marilyn C. Cornelis, Sean P. David, David A. Hinds, Eric Jorgenson, Jaakko Kaprio, Jerry A. Stitzel, Kari Stefansson, Thorgeir E. Thorgeirsson, Gonçalo Abecasis, Dajiang J. Liu, and Scott Vrieze. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*, 51(2):237–244, February 2019.

[106] Alexander Lorbert and Peter Ramadge. Descent Methods for Tuning Parameter Refinement. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 469–476, March 2010.

[107] Brenton Louie, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, and Peter Tarczy-Hornoch. Data integration and genomic medicine. *Journal of Biomedical Informatics*, 40(1):5–16, February 2007.

[108] Ganfeng Luo, Yanting Zhang, Pi Guo, Huanlin Ji, Yuejiao Xiao, and Ke Li. Global Patterns and Trends in Pancreatic Cancer Incidence: Age, Period, and Birth Cohort Analysis. *Pancreas*, 48(2):199–208, 2019.

[109] Shannon M. Lynch, Alina Vrieling, Jay H. Lubin, Peter Kraft, Julie B. Mendelsohn, Patricia Hartge, Federico Canzian, Emily Steplowski, Alan A. Arslan, Myron Gross, Kathy Helzlsouer, Eric J. Jacobs, Andrea LaCroix, Gloria Petersen, Wei Zheng, Demetrius Albanes, Laufey Amundadottir, Sheila A. Bingham, Paolo Boffetta, Marie-Christine Boutron-Ruault, Stephen J. Chanock, Sandra Clipp, Robert N. Hoover, Kevin Jacobs, Karen C. Johnson, Charles Kooperberg, Juhua Luo, Catherine Messina, Domenico Palli, Alpa V. Patel, Elio Riboli, Xiao-Ou Shu, Laudina Rodriguez Suarez, Gilles Thomas, Anne Tjønneland, Geoffrey S. Tobias, Elissa Tong, Dimitrios Trichopoulos, Jarmo Virtamo, Weimin Ye, Kai Yu, Anne Zeleniuch-Jacquette, H. Bas Bueno-de Mesquita, and Rachael Z. Stolzenberg-Solomon. Cigarette smoking and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort

consortium. *American Journal of Epidemiology*, 170(4):403–413, August 2009.

[110] Paige Maas, Myrto Barrdahl, Amit D. Joshi, Paul L. Auer, Mia M. Gaudet, Roger L. Milne, Fredrick R. Schumacher, William F. Anderson, David Check, Subham Chattopadhyay, Laura Baglietto, Christine D. Berg, Stephen J. Chanock, David G. Cox, Jonine D. Figueroa, Mitchell H. Gail, Barry I. Graubard, Christopher A. Haiman, Susan E. Hankinson, Robert N. Hoover, Claudine Isaacs, Laurence N. Kolonel, Loic Le Marchand, I.-Min Lee, Sara Lindström, Kim Overvad, Isabelle Romieu, Maria-Jose Sanchez, Melissa C. Southey, Daniel O. Stram, Rosario Tumino, Tyler J. VanderWeele, Walter C. Willett, Shumin Zhang, Julie E. Buring, Federico Canzian, Susan M. Gapstur, Brian E. Henderson, David J. Hunter, Graham G. Giles, Ross L. Prentice, Regina G. Ziegler, Peter Kraft, Montse Garcia-Closas, and Nilanjan Chatterjee. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states. 2(10):1295–1302.

[111] Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469–480, 2017.

[112] Charles F. Manski and Steven R. Lerman. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica*, 45(8):1977–1988, 1977.

[113] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, July 2010.

[114] P. Mass et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states. *JAMA Oncology*, 2(10):1295–1302, 2016.

[115] T. Mathew and K. Nordstrom. On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics*, 55(4):1221–3, 1999.

[116] Nana Matoba, Masato Akiyama, Kazuyoshi Ishigaki, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Shiro Ikegawa, Masashi Ikeda, Nakao Iwata, Makoto Hirata, Koichi Matsuda, Michiaki Kubo, Yukinori Okada, and Yoichiro Kamatani. GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nature Human Behaviour*, 3(5):471–477, 2019.

[117] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 2nd edition, 1989.

[118] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, Penny Soucy, Dylan Glubb, Asha Rostamianfar, Manjeet K. Bolla, Qin Wang, Jonathan Tyrer, Ed Dicks, Andrew Lee, Zhaoming Wang, Jamie Allen, Renske Keeman, Ursula Eilber, Juliet D. French, Xiao Qing Chen, Laura Fachal, Karen McCue, Amy E. McCart Reed, Maya Ghoussaini, Jason S. Carroll, Xia

Jiang, Hilary Finucane, Marcia Adams, Muriel A. Adank, Habibul Ahsan, Kristiina Aittomäki, Hoda Anton-Culver, Natalia N. Antonenkova, Volker Arndt, Kristan J. Aronson, Banu Arun, Paul L. Auer, François Bacot, Myrto Barrdahl, Caroline Baynes, Matthias W. Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V. Bogdanova, Stig E. Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Judith S. Brand, Hiltrud Brauch, Paul Brennan, Hermann Brenner, Louise Brinton, Per Broberg, Ian W. Brock, Annegien Broeks, Angela Brooks-Wilson, Sara Y. Brucker, Thomas Brüning, Barbara Burwinkel, Katja Butterbach, Qiuyin Cai, Hui Cai, Trinidad Caldés, Federico Canzian, Angel Carracedo, Brian D. Carter, Jose E. Castelao, Tsun L. Chan, Ting-Yuan David Cheng, Kee Seng Chia, Ji-Yeob Choi, Hans Christiansen, Christine L. Clarke, Margriet Collée, Don M. Conroy, Emilie Cordina-Duverger, Sten Cornelissen, David G. Cox, Angela Cox, Simon S. Cross, Julie M. Cunningham, Kamila Czene, Mary B. Daly, Peter Devilee, Kimberly F. Doheny, Thilo Dörk, Isabel dos Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M. Eccles, Arif B. Ekici, A. Heather Eliassen, Carolina Ellberg, Mingajeva Elvira, Christoph Engel, Mikael Eriksson, Peter A. Fasching, Jonine Figueroa, Dieter Flesch-Janys, Olivia Fletcher, Henrik Flyger, Lin Fritschi, Valerie Gaborieau, Marike Gabrielson, Manuela Gago-Dominguez, Yu-Tang Gao, Susan M. Gapstur, José A. García-Sáenz, Mia M. Gaudet, Vassilios Georgoulias, Graham G. Giles, Gord Glendon, Mark S. Goldberg, David E. Goldgar, Anna González-Neira, Grethe I. Grenaker Alnæs, Mervi Grip, Jacek

Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A. Haiman, Niclas Håkansson, Ute Hamann, Nathalie Hamel, Susan Hankinson, Patricia Harrington, Steven N. Hart, Jaana M. Hartikainen, Mikael Hartman, Alexander Hein, Jane Heyworth, Belynda Hicks, Peter Hillemanns, Dona N. Ho, Antoinette Hollestelle, Maartje J. Hooning, Robert N. Hoover, John L. Hopper, Ming-Feng Hou, Chia-Ni Hsiung, Guanmengqian Huang, Keith Humphreys, Junko Ishiguro, Hidemi Ito, Motoki Iwasaki, Hiroji Iwata, Anna Jakubowska, Wolfgang Janni, Esther M. John, Nichola Johnson, Kristine Jones, Michael Jones, Arja Jukkola-Vuorinen, Rudolf Kaaks, Maria Kabisch, Katarzyna Kaczmarek, Daehee Kang, Yoshio Kasuga, Michael J. Kerin, Sofia Khan, Elza Khusnutdinova, Johanna I. Kiiski, Sung-Won Kim, Julia A. Knight, Veli-Matti Kosma, Vessela N. Kristensen, Ute Krüger, Ava Kwong, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Min Hyuk Lee, Jong Won Lee, Chuen Neng Lee, Flavio Lejbkowicz, Jingmei Li, Jenna Lilyquist, Annika Lindblom, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Artitaya Lophatananon, Jan Lubinski, Craig Luccarini, Michael P. Lux, Edmond S. K. Ma, Robert J. MacInnis, Tom Maishman, Enes Makalic, Kathleen E. Malone, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, JoAnn E. Manson, Sara Margolin, Shivaani Mariapun, Maria Elena Martinez, Keitaro Matsuo, Dimitrios Mavroudis, James McKay, Catriona McLean, Hanne Meijers-Heijboer, Alfons Meindl, Primitiva Menéndez, Usha Menon, Jeffery Meyer, Hui Miao, Nicola Miller, Nur Aishah Mohd Taib, Kenneth Muir, Anna Marie Mulligan, Claire Mulot, Susan L. Neuhausen, Heli Nevanlinna, Patrick Neven, Sune F. Nielsen, Dong-Young Noh, Børge G. Nordestgaard, Aaron Norman, Olufunmilayo I. Olopade, Janet E. Olson, Håkan Olsson, Curtis Olswold, Nick Orr, V. Shane Pankratz, Sue K. Park, Tjoung-Won Park-Simon, Rachel Lloyd, Jose I. A. Perez, Paolo Peterlongo, Julian Peto, Kelly-Anne Phillips, Mila Pinchev, Dijana Plaseska-Karanfilska, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Elizabeth Pugh, Katri Pylkäs, Brigitte Rack, Paolo Radice, Nazneen Rahman, Gadi Rennert, Hedy S. Rennert, Valerie Rhenius, Atocha Romero, Jane Romm, Kathryn J. Ruddy, Thomas Rüdiger, Anja Rudolph, Matthias Ruebner, Emiel J. T. Rutgers, Emmanouil Saloustros, Dale P. Sandler, Suleeporn Sangrajrang, Elinor J. Sawyer, Daniel F. Schmidt, Rita K. Schmutzler, Andreas Schneeweiss, Minouk J. Schoemaker, Fredrick Schumacher, Peter Schürmann, Rodney J. Scott, Christopher Scott, Sheila Seal, Caroline Seynaeve, Mitul Shah, Priyanka Sharma, Chen-Yang Shen, Grace Sheng, Mark E. Sherman, Martha J. Shrubsole, Xiao-Ou Shu, Ann Smeets, Christof Sohn, Melissa C. Southey, John J. Spinelli, Christa Stegmaier, Sarah Stewart-Brown, Jennifer Stone, Daniel O. Stram, Harald Surowy, Anthony Swerdlow, Rulla Tamimi, Jack A. Taylor, Maria Tengström, Soo H. Teo, Mary Beth Terry, Daniel C. Tessier, Somchai Thanasitthichai, Kathrin Thöne, Rob A. E. M. Tollenaar, Ian Tomlinson, Ling Tong, Diana Torres, Thérèse Truong, Chiu-Chen Tseng, Shoichiro Tsugane, Hans-Ulrich Ulmer, Giske Ursin, Michael Untch, Celine Vachon, Christi J. van Asperen, David Van Den Berg, Ans M. W. van den Ouweland, Lizet van der Kolk, Rob B. van der Luijt, Daniel Vincent, Jason Vollenweider, Quinten Waisfisz, Shan Wang-Gohrke, Clarice R. Weinberg, Camilla Wendt, Alice S. Whittemore,

Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Anna H. Wu, Lucy Xia, Taiki Yamaji, Xiaohong R. Yang, Cheng Har Yip, Keun-Young Yoo, Jyh-Cherng Yu, Wei Zheng, Ying Zheng, Bin Zhu, Argyrios Ziogas, Elad Ziv, Sunil R. Lakhani, Antonis C. Antoniou, Arnaud Droit, Irene L. Andrulis, Christopher I. Amos, Fergus J. Couch, Paul D. P. Pharoah, Jenny Chang-Claude, Per Hall, David J. Hunter, Roger L. Milne, Montserrat García-Closas, Marjanka K. Schmidt, Stephen J. Chanock, Alison M. Dunning, Stacey L. Edwards, Gary D. Bader, Georgia Chenevix-Trench, Jacques Simard, Peter Kraft, and Douglas F. Easton. Association analysis identifies 65 new breast cancer risk loci. 551(7678):92–94.

[119] Stephen J Mooney, Daniel J Westreich, and Abdulrahman M El-Sayed. Epidemiology in the era of big data. 26(3):390–394.

[120] Bhramar Mukherjee, Jaeil Ahn, Stephen B. Gruber, and Nilanjan Chatterjee. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology*, 175(3):177–190, February 2012.

[121] Bhramar Mukherjee and Nilanjan Chatterjee. Exploiting Gene-Environment Independence for Analysis of Case-Control Studies: An Empirical Bayes-Type Shrinkage Estimator to Trade-Off between Bias and Efficiency. *Biometrics*, 64(3):685–694, September 2008.

[122] Mike A. Nalls, Nathan Pankratz, Christina M. Lill, Chuong B. Do, Dena G. Hernandez, Mohamad Saad, Anita L. DeStefano, Eleanna Kara,

Jose Bras, Manu Sharma, Claudia Schulte, Margaux F. Keller, Sampath Arepalli, Christopher Letson, Connor Edsall, Hreinn Stefansson, Xinmin Liu, Hannah Pliner, Joseph H. Lee, Rong Cheng, International Parkinson's Disease Genomics Consortium (IPDGC), Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI), 23andMe, GenePD, NeuroGenetics Research Consortium (NGRC), Hussman Institute of Human Genomics (HIHG), Ashkenazi Jewish Dataset Investigator, Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE), North American Brain Expression Consortium (NABEC), United Kingdom Brain Expression Consortium (UKBEC), Greek Parkinson's Disease Consortium, Alzheimer Genetic Analysis Group, M. Arfan Ikram, John P. A. Ioannidis, Georgios M. Hadjigeorgiou, Joshua C. Bis, Maria Martinez, Joel S. Perlmutter, Alison Goate, Karen Marder, Brian Fiske, Margaret Sutherland, Georgia Xiromerisiou, Richard H. Myers, Lorraine N. Clark, Kari Stefansson, John A. Hardy, Peter Heutink, Honglei Chen, Nicholas W. Wood, Henry Houlden, Haydeh Payami, Alexis Brice, William K. Scott, Thomas Gasser, Lars Bertram, Nicholas Eriksson, Tatiana Foroud, and Andrew B. Singleton. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics*, 46(9):989–993, September 2014.

[123] National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Reports of

the Surgeon General. Centers for Disease Control and Prevention (US), Atlanta (GA), 2014.

[124] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume Volume 4, pages 2111–2245. Elsevier, 1994.

[125] J. Neyman. Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association*, 33(201):101–116, March 1938.

[126] Leticia M. Nogueira, Christina C. Newton, Michael Pollak, Debra T. Silverman, Demetrius Albanes, Satu Männistö, Stephanie J. Weinstein, Eric J. Jacobs, and Rachael Z. Stolzenberg-Solomon. Serum C-peptide, Total and High Molecular Weight Adiponectin, and Pancreatic Cancer: Do Associations Differ by Smoking? *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 26(6):914–922, 2017.

[127] I. Olkin and A. Sampson. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, 54(1):317–322, 1998.

[128] B Pasaniuc and A.L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet.*, 18(2):117–127, 2017.

[129] Margaret Sullivan Pepe and Thomas R. Fleming. A Nonparametric Method for Dealing With Mismeasured Covariate Data. *Journal of the American Statistical Association*, 86(413):108–113, 1991.

[130] Gloria M. Petersen, Laufey Amundadottir, Charles S. Fuchs, Peter Kraft, Rachael Z. Stolzenberg-Solomon, Kevin B. Jacobs, Alan A. Arslan, H. Bas Bueno-de Mesquita, Steven Gallinger, Myron Gross, Kathy Helzlsouer, Elizabeth A. Holly, Eric J. Jacobs, Alison P. Klein, Andrea LaCroix, Donghui Li, Margaret T. Mandelson, Sara H. Olson, Harvey A. Risch, Wei Zheng, Demetrius Albanes, William R. Bamlet, Christine D. Berg, Marie-Christine Boutron-Ruault, Julie E. Buring, Paige M. Bracci, Federico Canzian, Sandra Clipp, Michelle Cotterchio, Mariza de Andrade, Eric J. Duell, J. Michael Gaziano, Edward L. Giovannucci, Michael Goggins, Göran Hallmans, Susan E. Hankinson, Manal Hassan, Barbara Howard, David J. Hunter, Amy Hutchinson, Mazda Jenab, Rudolf Kaaks, Charles Kooperberg, Vittorio Krogh, Robert C. Kurtz, Shannon M. Lynch, Robert R. McWilliams, Julie B. Mendelsohn, Dominique S. Michaud, Hemang Parikh, Alpa V. Patel, Petra H. M. Peeters, Aleksandar Rajkovic, Elio Riboli, Laudina Rodriguez, Daniela Seminara, Xiao-Ou Shu, Gilles Thomas, Anne Tjønneland, Geoffrey S. Tobias, Dimitrios Trichopoulos, Stephen K. Van Den Eeden, Jarmo Virtamo, Jean Wactawski-Wende, Zhaoming Wang, Brian M. Wolpin, Herbert Yu, Kai Yu, Anne Zeleniuch-Jacquotte, Joseph F. Fraumeni, Robert N. Hoover, Patricia Hartge, and Stephen J. Chanock. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*, 42(3):224–228, March 2010.

[131] Danny Pfeffermann. The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review / Revue Internationale de*

Statistique, 61(2):317–337, 1993.

[132] Andrew Pickles, Graham Dunn, and José Luis Vázquez-Barquero. Screening for stratification in two-phase ('two- stage') epidemiological surveys. *Statistical Methods in Medical Research*, 4(1):73–89, March 1995.

[133] Walter W. Piegorsch, Clarice R. Weinberg, and Jack A. Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2):153–162, 1994.

[134] R. L. Prentice. A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika*, 73(1):1–11, 1986.

[135] Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994.

[136] Jing Qin. Combining parametric and empirical likelihoods. *Biometrika*, 87(2):484–490, 2000.

[137] Jing Qin, Han Zhang, Pengfei Li, Demetrius Albanes, and Kai Yu. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102(1):169–180, March 2015.

[138] Hazhir Rahmandad, Mohammad S. Jalali, and Kamran Paynabar. A flexible method for aggregation of prior statistical findings. *PloS one*, 12(4):e0175111, 2017.

[139] Prashanth Rawla, Tagore Sunkara, and Vinaya Gaduputi. Epidemiology

of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World Journal of Oncology*, 10(1):10–27, February 2019.

[140] J. Ritz, E. Demidenko, and D. Spiegelman. Multivariate meta-analysis for data consortia, individual patient meta-analysis, and pooling projects. *Journal of Statistical Planning and Inference*, 138(7):1919–1933, 2008.

[141] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994.

[142] T.J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

[143] Walter Rudin. *Principles of Mathematical Analysis.* McGraw Hill, 3rd edition, 1976.

[144] Takumi Saegusa and Jon A. Wellner. WEIGHTED LIKELIHOOD ESTIMATION UNDER TWO-PHASE SAMPLING. *The Annals of Statistics*, 41(1):269–295, 2013.

[145] Daniel J. Schaid, Gregory D. Jenkins, James N. Ingle, and Richard M. Weinshilboum. Two-Phase Designs to Follow-Up Genome-Wide Association Signals With DNA Resequencing Studies. *Genetic epidemiology*, 37(3), April 2013.

[146] A. J. Scott and C. J. Wild. Fitting regression models to case-control data by maximum likelihood. 84(1):57–71.

[147] A.J. Scott and C.J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):705–717, 1997.

[148] Alastair Scott and Christopher Wild. Calculating efficient semiparametric estimators for a broad class of missing-data problems. *Festschrift for Tarmo Pukkila on his 60th Birthday*, January 2006.

[149] David B. Searls. Data integration: challenges for drug discovery. 4(1):45–58.

[150] Rebecca L. Siegel, Eric J. Jacobs, Christina C. Newton, Diane Feskanich, Neal D. Freedman, Ross L. Prentice, and Ahmedin Jemal. Deaths Due to Cigarette Smoking for 12 Smoking-Related Cancers in the United States. *JAMA Internal Medicine*, 175(9):1574–1576, September 2015.

[151] Minsun Song, William Wheeler, Neil E. Caporaso, Maria Teresa Landi, and Nilanjan Chatterjee. Using imputed genotype data in the joint score tests for genetic association and gene-environment interactions in case-control studies. *Genetic epidemiology*, 42(2):146–155, March 2018.

[152] Rubina Tabassum, Ganesh Chauhan, Om Prakash Dwivedi, Anubha Mahajan, Alok Jaiswal, Ismeet Kaur, Khushdeep Bandesh, Tejbir Singh, Benan John Mathai, Yogesh Pandey, Manickam Chidambaram, Amitabh Sharma, Sreenivas Chavali, Shantanu Sengupta, Lakshmi Ramakrishnan, Pradeep Venkatesh, Sanjay K. Aggarwal, Saurabh Ghosh, Dorairaj Prabhakaran, Reddy K. Srinath, Madhukar Saxena, Monisha Banerjee, Sandeep Mathur, Anil Bhansali, Viral N. Shah, Sri Venkata Madhu, Raman K. Marwaha, Analabha Basu, Vinod Scaria, Mark I. McCarthy,

Radha Venkatesan, Viswanathan Mohan, Nikhil Tandon, and Dwaipayan Bharadwaj. Genome-Wide Association Study for Type 2 Diabetes in Indians Identifies a New Susceptibility Locus at 2q21. *Diabetes*, 62(3):977–986, March 2013.

[153] Hongwei Tang, Peng Wei, Eric J. Duell, Harvey A. Risch, Sara H. Olson, H. Bas Bueno-de Mesquita, Steven Gallinger, Elizabeth A. Holly, Gloria Petersen, Paige M. Bracci, Robert R. McWilliams, Mazda Jenab, Elio Riboli, Anne Tjønneland, Marie Christine Boutron-Ruault, Rudolph Kaaks, Dimitrios Trichopoulos, Salvatore Panico, Malin Sund, Petra H. M. Peeters, Kay-Tee Khaw, Christopher I. Amos, and Donghui Li. Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene- and pathway-based interaction analysis of GWAS data. *Carcinogenesis*, 35(5):1039–1045, May 2014.

[154] Duncan C. Thomas, Zhao Yang, and Fan Yang. Two-phase and family-based designs for next-generation sequencing studies. *Frontiers in Genetics*, 4, 2013.

[155] Hui-Jen Tsai and Jeffrey S. Chang. Environmental Risk Factors of Pancreatic Cancer. *Journal of Clinical Medicine*, 8(9), September 2019.

[156] H.C. van Houwelingen, L.R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002.

[157] S. Wacholder and R.J. Carroll. The partial questionnaire design for case-control studies. *Statistics in Medicine*, 13(5-7):623–634, 1994.

[158] Sholom Wacholder, Raymond J. Carroll, David Pee, and Mitchell H. Gail. The partial questionnaire design for case-control studies. *Statistics in Medicine*, 13(5-7):623–634, 1994.

[159] Sholom Wacholder, Mitchell H. Gail, David Pee, and Ron Brookmeyer. Alternative Variance and Efficiency Calculations for the Case-Cohort Design. *Biometrika*, 76(1):117–123, 1989.

[160] Fei Wang, Peter X.-K. Song, and Lu Wang. Merging multiple longitudinal studies with study-specific missing covariates: A joint estimating function approach. *Biometrics*, 71(4):929–940, 2015.

[161] Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, 8(1):1–11, November 2017.

[162] A. Whittemore. Multistage sampling designs and estimating equations. *J. R. Statist. Soc. B*, 59(3):589–602, 1997.

[163] Alice S. Whittemore and Jerry Halpern. Multi-Stage Sampling in Genetic Epidemiology. *Statistics in Medicine*, 16(2):153–167, 1997.

[164] C. J. Wild. Fitting prospective regression models to case-control data. *Biometrika*, 78(4):705–717, December 1991.

[165] Uwe A. Wittel, Ulrich T. Hopt, and Surinder K. Batra. Cigarette smoke-induced pancreatic damage—experimental data. *Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie*, 393(4), July 2008.

[166] Brian M. Wolpin, Cosmeri Rizzato, Peter Kraft, Charles Kooperberg, Gloria M. Petersen, Zhaoming Wang, Alan A. Arslan, Laura Beane-Freeman, Paige M. Bracci, Julie Buring, Federico Canzian, Eric J. Duell, Steven Gallinger, Graham G. Giles, Gary E. Goodman, Phyllis J. Goodman, Eric J. Jacobs, Aruna Kamineni, Alison P. Klein, Laurence N. Kolonel, Matthew H. Kulke, Donghui Li, Núria Malats, Sara H. Olson, Harvey A. Risch, Howard D. Sesso, Kala Visvanathan, Emily White, Wei Zheng, Christian C. Abnet, Demetrius Albanes, Gabriella Andreotti, Melissa A. Austin, Richard Barfield, Daniela Basso, Sonja I. Berndt, Marie-Christine Boutron-Ruault, Michelle Brotzman, Markus W. Büchler, H. Bas Bueno-de Mesquita, Peter Bugert, Laurie Burdette, Daniele Campa, Neil E. Caporaso, Gabriele Capurso, Charles Chung, Michelle Cotterchio, Eithne Costello, Joanne Elena, Niccola Funel, J. Michael Gaziano, Nathalia A. Giese, Edward L. Giovannucci, Michael Goggins, Megan J. Gorman, Myron Gross, Christopher A. Haiman, Manal Hassan, Kathy J. Helzlsouer, Brian E. Henderson, Elizabeth A. Holly, Nan Hu, David J. Hunter, Federico Innocenti, Mazda Jenab, Rudolf Kaaks, Timothy J. Key, Kay-Tee Khaw, Eric A. Klein, Manolis Kogevinas, Vittorio Krogh, Juozas Kupcinskas, Robert C. Kurtz, Andrea LaCroix, Maria T. Landi, Stefano Landi, Loic Le Marchand, Andrea Mambrini, Satu Mannisto, Roger L. Milne, Yusuke Nakamura, Ann L. Oberg, Kouros Owzar, Alpa V. Patel, Petra H. M. Peeters, Ulrike Peters, Raffaele Pezzilli, Ada Piepoli, Miquel Porta, Francisco X. Real, Elio Riboli, Nathaniel Rothman, Aldo Scarpa, Xiao-Ou Shu, Debra T. Silverman, Pavel Soucek, Malin Sund, Renata Talar-Wojnarowska, Philip R. Taylor, George E. Theodoropoulos, Mark

Thornquist, Anne Tjønneland, Geoffrey S. Tobias, Dimitrios Trichopoulos, Pavel Vodicka, Jean Wactawski-Wende, Nicolas Wentzensen, Chen Wu, Herbert Yu, Kai Yu, Anne Zeleniuch-Jacquotte, Robert Hoover, Patricia Hartge, Charles Fuchs, Stephen J. Chanock, Rachael S. Stolzenberg-Solomon, and Laufey T. Amundadottir. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nature Genetics*, 46(9):994–1000, September 2014.

[167] J. Yang and M. D. Li. Converging findings from linkage and association analyses on susceptibility genes for smoking and other addictions. *Molecular Psychiatry*, 21(8):992–1008, 2016.

[168] Jian Yang, Teresa Ferreira, Andrew P. Morris, Sarah E. Medland, Pamela AF Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, Michael N. Weedon, and Ruth J. Loos. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369, 2012.

[169] Matthew B. Yurgelun, Anu B. Chittenden, Vicente Morales-Oyarvide, Douglas A. Rubinson, Richard F. Dunne, Margaret M. Kozak, Zhi Rong Qian, Marisa W. Welch, Lauren K. Brais, Annacarolina Da Silva, Justin L. Bui, Chen Yuan, Tingting Li, Wanwan Li, Atsuhiro Masuda, Mancang Gu, Andrea J. Bullock, Daniel T. Chang, Thomas E. Clancy, David C. Linehan, Jennifer J. Findeis-Hosey, Leona A. Doyle, Aaron R. Thorner, Matthew D. Ducar, Bruce M. Wollison, Natalia Khalaf, Kimberly Perez, Sapna Syngal, Andrew J. Aguirre, William C. Hahn, Matthew L. Meyerson, Charles S. Fuchs, Shuji Ogino, Jason L. Hornick, Aram F. Hezel,

Albert C. Koong, Jonathan A. Nowak, and Brian M. Wolpin. Germline cancer susceptibility gene variants, somatic second hits, and survival outcomes in patients with resected pancreatic cancer. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 21(1):213–223, 2019.

[170] Mingfeng Zhang, Zhaoming Wang, Ofure Obazee, Jinping Jia, Erica J. Childs, Jason Hoskins, Gisella Figlioli, Evelina Mocci, Irene Collins, Charles C. Chung, Christopher Hautman, Alan A. Arslan, Laura Beane-Freeman, Paige M. Bracci, Julie Buring, Eric J. Duell, Steven Gallinger, Graham G. Giles, Gary E. Goodman, Phyllis J. Goodman, Aruna Kamineni, Laurence N. Kolonel, Matthew H. Kulke, Núria Malats, Sara H. Olson, Howard D. Sesso, Kala Visvanathan, Emily White, Wei Zheng, Christian C. Abnet, Demetrius Albanes, Gabriella Andreotti, Lauren Brais, H. Bas Bueno-de Mesquita, Daniela Basso, Sonja I. Berndt, Marie-Christine Boutron-Ruault, Maarten F. Bijlsma, Hermann Brenner, Laurie Burdette, Daniele Campa, Neil E. Caporaso, Gabriele Capurso, Giulia Martina Cavestro, Michelle Cotterchio, Eithne Costello, Joanne Elena, Ugo Boggi, J. Michael Gaziano, Maria Gazouli, Edward L. Giovannucci, Michael Goggins, Myron Gross, Christopher A. Haiman, Manal Hassan, Kathy J. Helzlsouer, Nan Hu, David J. Hunter, Elzbieta Iskierka-Jazdzewska, Mazda Jenab, Rudolf Kaaks, Timothy J. Key, Kay-Tee Khaw, Eric A. Klein, Manolis Kogevinas, Vittorio Krogh, Juozas Kupcinskas, Robert C. Kurtz, Maria T. Landi, Stefano Landi, Loic Le Marchand, Andrea Mambrini, Satu Mannisto, Roger L. Milne, Rachel E. Neale,

Ann L. Oberg, Salvatore Panico, Alpa V. Patel, Petra H. M. Peeters, Ulrike Peters, Raffaele Pezzilli, Miquel Porta, Mark Purdue, J. Ramón Quiros, Elio Riboli, Nathaniel Rothman, Aldo Scarpa, Ghislaine Scelo, Xiao-Ou Shu, Debra T. Silverman, Pavel Soucek, Oliver Strobel, Malin Sund, Ewa Małecka-Panas, Philip R. Taylor, Francesca Tavano, Ruth C. Travis, Mark Thornquist, Anne Tjønneland, Geoffrey S. Tobias, Dimitrios Trichopoulos, Yogesh Vashist, Pavel Vodicka, Jean Wactawski-Wende, Nicolas Wentzensen, Herbert Yu, Kai Yu, Anne Zeleniuch-Jacquotte, Charles Kooperberg, Harvey A. Risch, Eric J. Jacobs, Donghui Li, Charles Fuchs, Robert Hoover, Patricia Hartge, Stephen J. Chanock, Gloria M. Petersen, Rachael S. Stolzenberg-Solomon, Brian M. Wolpin, Peter Kraft, Alison P. Klein, Federico Canzian, and Laufey T. Amundadottir. Three new pancreatic cancer susceptibility signals identified on chromosomes 1q32.1, 5p15.33 and 8q24.21. *Oncotarget*, 7(41):66328–66343, October 2016.

[171] Haibo Zhou, Rui Song, Yuanshan Wu, and Jing Qin. Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics*, 67(1):194–202, March 2011.

[172] Qingning Zhou, Jianwen Cai, and Haibo Zhou. Semiparametric inference for a two-stage outcome-dependent sampling design with interval-censored failure time data. *Lifetime Data Analysis*, January 2019.

[173] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R. Robinson, Joseph E. Powell, Grant W. Montgomery, Michael E. Goddard,

Naomi R. Wray, Peter M. Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. 48(5):481–487.

Appendix A

Chapter 1

A.1 Asymptotic Equivalence of GENMETA Estimator and Simple Meta-Analysis Estimator When All the Reduced Models Are the Same to the Maximal Model

When all the reduced models are the same to the maximal model, it follows $\theta_k^* = \beta^*$, $X_{A_k} = X$ and $g_k = f$ for $k = 1, 2, \ldots, K$. Then, for each k, $u_k(X; \beta^*, \theta_k^*) = u_k(X; \beta^*, \beta^*) = \int s_k(y \mid X_{A_k}; \beta^*) f(y \mid X; \beta^*) dy = 0$. By the definition of Δ, we have $\Delta = 0$. On the other hand, assuming $E_{Y|X}\{\nabla_{\theta_k} s_k(\theta_k^*)\} = \nabla_{\theta_k} E_{Y|X}\{s_k(\theta_k^*)\}$ with $s_k(\theta_k^*) = s_k(Y \mid X_{A_k}; \theta_k^*)$, it follows $\Lambda_k = (1/c_k) I(\theta_k^*)$, where $I(\theta_k^*)$ is the Fisher's information matrix of g_k or f. Then, the optimal C is

$$C_{\text{opt}} = \Lambda^{-1} = \text{diag}(c_1\Sigma, \ldots, c_K\Sigma),$$

where $\Sigma = I(\theta_k^*)^{-1}$. Denote as $\hat{C}_{\text{opt}}$ a consistent estimator of C_{opt}. Then, the GENMETA estimator with $\hat{C}_{\text{opt}}$ is

$$\hat{\beta}_{\text{opt}} = \text{argmin}_{\beta} U_n^T(\beta, \hat{\theta}) \hat{C}_{\text{opt}} U_n(\beta, \hat{\theta}).$$

Under regularity conditions similar to those in Theorem 1, $\hat{\beta}_{\text{opt}} \to \beta^*$ in probability. By Mean Value Theorem,

$$U_n(\hat{\beta}_{\text{opt}}, \hat{\theta}) = U_n(\beta^*, \hat{\theta}) + G_n(\bar{\beta}, \hat{\theta})(\hat{\beta}_{\text{opt}} - \beta^*), \tag{A.1}$$

where $\bar{\beta}$ is the mean value and $G_n(\bar{\beta}, \hat{\theta}) = \partial U_n(\beta, \hat{\theta})/\partial\beta \mid_{\beta=\bar{\beta}}$. By the first order condition, $\hat{\beta}_{\text{opt}}$ satisfies $G_n^T(\hat{\beta}_{\text{opt}}, \hat{\theta}) \hat{C}_{\text{opt}} U_n(\hat{\beta}_{\text{opt}}, \hat{\theta}) = 0$. Left-multiplying (A.1) by $G_n^T(\hat{\beta}_{\text{opt}}, \hat{\theta}) \hat{C}_{\text{opt}}$, it follows

$$\hat{\beta}_{\text{opt}} - \beta^* = -\{G_n^T(\hat{\beta}_{\text{opt}}, \hat{\theta}) \hat{C}_{\text{opt}} G_n(\bar{\beta}, \hat{\theta})\}^{-1} \{G_n^T(\hat{\beta}_{\text{opt}}, \hat{\theta}) \hat{C}_{\text{opt}} U_n(\beta^*, \hat{\theta})\} \tag{A.2}$$

Also,

$$G_n(\hat{\beta}_{\text{opt}}, \hat{\theta}) = \frac{\partial}{\partial\beta} U_n(\beta, \hat{\theta}) \mid_{\beta=\hat{\beta}_{\text{opt}}} = \begin{pmatrix} \frac{\partial}{\partial\beta} u_1(\beta, \hat{\theta}_1) \mid_{\beta=\hat{\beta}_{\text{opt}}} \\ \vdots \\ \frac{\partial}{\partial\beta} u_K(\beta, \hat{\theta}_K) \mid_{\beta=\hat{\beta}_{\text{opt}}} \end{pmatrix}.$$

Under regularity conditions similar to those in Theorem 1, $\partial u_k(\beta, \hat{\theta}_k)/\partial\beta \mid_{\beta=\hat{\beta}_{\text{opt}}} = \Sigma^{-1} + o_p(1)$ for each k. Then,

$$G_n(\hat{\beta}_{\text{opt}}, \hat{\theta}) = \begin{pmatrix} \Sigma^{-1} \\ \vdots \\ \Sigma^{-1} \end{pmatrix} + o_p(1). \tag{A.3}$$

Similarly,

$$G_n(\bar{\beta}, \hat{\theta}) = \begin{pmatrix} \Sigma^{-1} \\ \vdots \\ \Sigma^{-1} \end{pmatrix} + o_p(1). \tag{A.4}$$

On the other hand, under regularity conditions similar to those in Theorem 1, $u_k(\beta^*, \hat{\theta}_k) = -\Sigma^{-1}(\hat{\theta}_k - \beta)^* + o_p(1/n^{1/2})$. Then,

$$U_n(\beta^*, \hat{\theta}) = - \begin{pmatrix} \Sigma^{-1}(\hat{\theta}_1 - \beta^*) \\ \vdots \\ \Sigma^{-1}(\hat{\theta}_K - \beta^*) \end{pmatrix} + o_p(1/n^{1/2}). \tag{A.5}$$

Hence, by (A.2), (A.3), (A.4), (A.5) and Slutsky's theorem,

$$\hat{\beta}_{\text{opt}} - \beta^* = \Big(\sum_{k=1}^{K} c_k\Big)^{-1} \Big\{\sum_{k=1}^{K} c_k(\hat{\theta}_k - \beta^*)\Big\} + o_p(1/n^{1/2}). \tag{A.6}$$

On the other hand,

$$\begin{aligned} \hat{\beta}_{\text{meta}} - \beta^* &= \Big\{\sum_{k=1}^{K} \Big(\frac{\hat{\Sigma}_k}{n_k}\Big)^{-1}\Big\}^{-1} \Big\{\sum_{k=1}^{K} \Big(\frac{\hat{\Sigma}_k}{n_k}\Big)^{-1} \hat{\theta}_k\Big\} - \beta^* \\ &= \Big(\sum_{k=1}^{K} c_k\Big)^{-1} \Big\{\sum_{k=1}^{K} c_k(\hat{\theta}_k - \beta^*)\Big\} + o_p(1/n^{1/2}). \end{aligned} \tag{A.7}$$

Therefore, by (A.6) and (A.7), $\hat{\beta}_{\text{opt}} = \hat{\beta}_{\text{meta}} + o_p(1/n^{1/2})$.

A.2 Newton-Raphson's Method and Iteratively Reweighted Least Squares Algorithm

In this section we provide a derivation of the Newton-Raphson's method for GENMETA with generalized linear models. As in Section 2.3, we assume that the maximal and reduced models belong to the class of GLM [117]. Specifically, assume the densities of $Y \mid X$ and $Y \mid X_{A_k}$ are of the forms

$$f(y \mid x; \beta, \phi) = \exp(\{1/a(\phi)\}(yh(x^T\beta) - b\{h(x^T\beta)\}) + c(y;\phi)),$$

and

$$g_k(y \mid x_{A_k}; \theta_k) = \exp(\{1/a(\phi_k)\}(yh(x_{A_k}^T\theta_k) - b\{h(x_{A_k}^T\theta_k)\}) + c(y;\phi_k)),$$

respectively, where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, $h(\cdot) = b'^{-1}(g^{-1}(\cdot))$, g is a monotone and differentiable link function, and ϕ and ϕ_k are the dispersion parameters of the maximal and the kth reduced models, respectively. Recall that we assume the maximal and the reduced models have the same link function g. However, both the GENMETA and the Newton-Raphson's method are flexible to allow the maximal and the reduced models to have different link functions. We also assume $X = \cup_{k=1}^{K} X_{A_k}$, where the vectors of the covariates are viewed as sets without confusion. Denote the dimensions of θ_k and β as d_k and p, respectively. Assume $d = \sum_{k=1}^{K} d_k \geq p$ since the parameters of the maximal model will not be identifiable if $d < p$.

A.2.1 Case I : ϕ and ϕ_k's are known.

The log-likelihood of g_k is

$$l_k(y \mid x_{A_k}; \theta_k) = \{1/a(\phi_k)\}(yh(x_{A_k}^T\theta_k) - b\{h(x_{A_k}^T\theta_k)\}) + c(y; \phi_k).$$

Then, the score function is

$$s_k(y \mid x_{A_k}; \theta_k) = \{1/a(\phi_k)\}\{y - g^{-1}(x_{A_k}^T\theta_k)\}h'(x_{A_k}^T\theta_k)x_{A_k}.$$

Then,

$$u_k(x; \beta, \theta_k) = E_{Y|X}s_k\{(y \mid x_{A_k}; \theta_k)\} = \{1/a(\phi_k)\}\{g^{-1}(x^T\beta) - g^{-1}(x_{A_k}^T\theta_k)\}h'(x_{A_k}^T\theta_k)x_{A_k}.$$

Thus, the vector of empirical moment functions for β is

$$U_n(\beta) = P_n \begin{pmatrix} u_k(X; \beta, \hat{\theta}_k) \\ u_k(X; \beta, \hat{\theta}_k) \\ \vdots \\ u_k(X; \beta, \hat{\theta}_k) \end{pmatrix},$$

where P_n is the empirical measure with respect to the reference sample.

Let $Q_n(\beta) = U_n^T(\beta)CU_n(\beta)$ where C is a $d \times d$ positive definite matrix. The goal is to find the minimizer of $Q_n(\beta)$. Its equivalent to solving the equation

$$D_n(\beta) = 0,$$

where $D_n(\beta) = G_n^T(\beta)CU_n(\beta)$ and $G_n(\beta) = \partial U_n(\beta)/\partial\beta$ is a $d \times p$ matrix. Then, the tth iteration step for the Newton-Raphson's method is

$$\beta^{(t+1)} = \beta^{(t)} - J_n(\beta^{(t)})^{-1}D_n(\beta^{(t)}), \tag{A.8}$$

where $J_n(\beta) = \partial D_n(\beta)/\partial\beta$ is a $p \times p$ matrix.

Next, we write $D_n(\beta)$ in a matrix form. The matrix form of $G_n(\beta)$ is

$$G_n(\beta) = P_n \begin{pmatrix} [a(\phi_1)g'\{g^{-1}(X^T\beta)\}]^{-1}h'(X_{A_1}^T\hat{\theta}_1)X_{A_1}X^T \\ \vdots \\ [a(\phi_K)g'\{g^{-1}(X^T\beta)\}]^{-1}h'(X_{A_K}^T\hat{\theta}_K)X_{A_K}X^T \end{pmatrix} = (1/n)X_{A_{diag}}^T W X_{rbind},$$

where $X_{\text{rbind}} = 1 \otimes X$ and $X_{(n\times p)}$ is the reference data matrix; $X_{A_{\text{diag}}} = \text{diag}(X_{A_1}, \ldots, X_{A_K})$ and $X_{A_k(n\times d_k)}$ is the reference data matrix for the kth study; $W = \text{diag}(W_1, \ldots, W_K)$, $W_k = \text{diag}(w_{k1}, \ldots, w_{kn})$, $w_{ki} = [a(\phi_k)g'\{g^{-1}(X_i^T\beta)\}]^{-1}h'(X_{A_k,i}^T\hat{\theta}_k)$ for $k = 1, \ldots, K$, $i = 1, \ldots, n$ and i, and X_i^T and $X_{A_k,i}^T$ are the ith rows of X and X_{A_k}, respectively. Similarly, the matrix form of $U_n(\beta)$ is $U_n(\beta) = (1/n)X_{A_{diag}}^T r$, where $r = (r_1, \ldots, r_K)^T$, $r_k = (r_{k1}, \ldots, r_{kn})^T$ and $r_{ki} = \{1/a(\phi_k)\}\{g^{-1}(X_i^T\beta) - g^{-1}(X_{A_k,i}^T\hat{\theta}_{A_k,i})\}h'(X_{A_k,i}^T\hat{\theta}_{A_k,i})$ for each k and i. Thus, the matrix form of $D_n(\beta)$ is

$$D_n(\beta) = (1/n^2)X_{rbind}^T W X_{A_{diag}} C X_{A_{diag}}^T r. \tag{A.9}$$

Next, we write $J_n(\beta)$ in a matrix form. Let $G_n(\beta)$ be partitioned by columns as $G_n(\beta) = (G_{n,1}(\beta), \ldots, G_{n,p}(\beta))$, where $G_{n,j}(\beta)$ is a $d \times 1$ column vector for $j = 1, \ldots, p$. Then,

$$\begin{aligned} J_n(\beta) &= \frac{\partial}{\partial \beta} D_n(\beta) = \frac{\partial}{\partial \beta} G_n^T(\beta) C U_n(\beta) \\ &= \begin{pmatrix} \frac{\partial}{\partial \beta} G_{n,1}^T(\beta) C U_n(\beta) \\ \vdots \\ \frac{\partial}{\partial \beta} G_{n,p}^T(\beta) C U_n(\beta) \end{pmatrix} = G_n^T(\beta) C G_n(\beta) + \begin{pmatrix} U_n^T(\beta) C \frac{\partial}{\partial \beta} G_{n,1}(\beta) \\ \vdots \\ U_n^T(\beta) C \frac{\partial}{\partial \beta} G_{n,p}(\beta) \end{pmatrix}. \end{aligned} \tag{A.10}$$

Then, the matrix form of the first summand is $(1/n^2) X_{rbind}^T W X_{A_{diag}} C X_{A_{diag}}^T W X_{rbind}$. The jth row of the second summand is $r^T X_{A_{diag}} C \partial G_{n,j}(\beta)/\partial \beta$. Note that

$$\frac{\partial}{\partial \beta} G_{n,j}(\beta) = (1/n) X_{A_{diag}}^T L X_{j_{diag}}^* X_{rbind},$$

where $L = \text{diag}(L_1, \ldots, L_K)$, $L_k = \text{diag}(l_{k1}, \ldots, l_{kn})$ and, for each k and i,

$$l_{ki} = -g''\{g^{-1}(X_i^T \beta)\}/(a(\phi_k)[g'\{g^{-1}(X_i^T \beta)\}]^3 h'(X_{A_k,i}^T \hat{\theta}_k));$$

$X_{j_{diag}}^* = \text{diag}(X_{j_{diag}}, \ldots, X_{j_{diag}})$ with K diagonal blocks and $X_{j_{diag}} = \text{diag}(X_{1j}, \ldots, X_{nj})$ for $j = 1, \ldots, p$. Then, for each j, the matrix form of $U_n^T(\beta) C \partial G_{n,j}(\beta)/\partial \beta$ is

$$(1/n^2) r^T X_{A_{diag}} C X_{A_{diag}}^T L X_{j_{diag}}^* X_{rbind}.$$

Then, the second summand of (A.10) can be rewritten as $(1/n^2) X_{rbind}^T V X_{rbind}$, where $V = diag(v_1, \ldots, v_{nK})$ and v_i is the ith element of the row vector $r^T X_{A_{diag}} C X_{A_{diag}}^T L$.

Thus,

$$J_n(\beta) = (1/n^2)X_{rbind}^T(WX_{A_{diag}}CX_{A_{diag}}^TW\ +\ V)X_{rbind} = (1/n^2)X_{\text{rbind}}^TW^*X_{\text{rbind}}. \tag{A.11}$$

where $W^* = WX_{A_{diag}}CX_{A_{diag}}^TW + V$.

Therefore, plugging (A.9) and (A.11) in (A.8), we get the following tth iteration step

$$\beta^{(t+1)} = \beta^{(t)} - (X_{\text{rbind}}^TW^*X_{\text{rbind}})^{-1}X_{\text{rbind}}^TWX_{A_{\text{diag}}}CX_{A_{\text{diag}}}^Tr,$$

which can be seen as the tth step of an iteratively reweighted least squares algorithm.

A.2.2 Case II : ϕ and ϕ_k's are unknown.

When ϕ and ϕ_k's are unknown, we propose to first obtain the GENMETA estimator $\hat{\beta}$ of $\beta^\star$ as above with $\phi_k's$ replaced by $\hat{\phi}_k$'s. Next, let us consider the estimation of $\phi^\star$, the true value of ϕ. For the kth reduced model, we have an additional score function with respect to ϕ_k, which is

$$s_k(y \mid x_{A_k};\theta_k,\phi_k) = -\frac{a'(\phi_k)}{a^2(\phi_k)}(yh(x_{A_k}^T\theta_k) - b\{h(x_{A_k}^T\theta_k)\}) + c'(y;\phi_k),$$

where $c'(y;\phi_k)$ is the derivative of $c(y;\phi_k)$ with respect to ϕ_k. Then, we obtain

$$u_k(X;\beta,\phi,\theta_k,\phi_k) = -\frac{a'(\phi_k)}{a^2(\phi_k)}(g^{-1}(X^T\beta)h(X_{A_k}^T\theta_k) - b\{h(X_{A_k}^T\theta_k)\}) + q_k(X;\beta,\phi,\phi_k),$$

where $q_k = E_{Y|X}(c'(Y, \phi_k))$. The distribution of $Y \mid X$ depends on β and ϕ so that q_k also depends on them. Then, the empirical moment vector for ϕ is

$$U_n(\phi) = P_n(u_1(X; \hat{\beta}, \phi, \hat{\theta}_1, \hat{\phi}_1)^T, \ldots, u_K(X; \hat{\beta}, \phi, \hat{\theta}_K, \hat{\phi}_K)^T)^T.$$

We propose to estimate $\phi^\star$ in the GMM framework. Thus, we need to compute the minimizer of $U_n(\phi)^T C U_n(\phi)$, where C is a known weighting matrix. As before, we use the Newton-Raphson's method and it can be written as

$$\phi^{(t+1)} = \phi^{(t)} - J_n^{-1}(\phi^{(t)}) D_n(\phi^{(t)}), \tag{A.12}$$

where

$$J_n(\phi) = U_n^T(\phi) C \frac{d^2}{d\phi^2} q_n(\phi) + (\frac{d}{d\phi} q_n(\phi))^T C \frac{d}{d\phi} q_n(\phi),$$

$D_n(\phi) = U_n^T(\phi^{(t)}) C dq_n(\phi)/d\phi$ and $q_n(\phi) = P_n(q_1(X; \hat{\beta}, \phi, \hat{\phi}_1), \ldots, q_K(X; \hat{\beta}, \phi, \hat{\phi}_K))^T$.

Thus, when ϕ and ϕ_k's are unknown, we first choose initial estimates $\beta^{(0)}$ and $\phi^{(0)}$. Then, we get the GENMETA estimator $\hat{\beta}$ by using equation (A.8) until a stopping rule is reached. Subsequently, $\phi^{(0)}$, $\hat{\beta}$ and the study estimates are plugged in equation (A.12) and the process is repeated until a stopping rule is reached to get the GENMETA estimator of ϕ^*. In each Newton-Raphson's step, the weighting matrix C is estimated by the estimates from the previous step.

If the estimates of the study dispersion parameters, ϕ_k's, are not provided directly, but the the outcomes are standardized ($\text{var}(Y) = 1$), we can obtain them through the following relation based on conditional variance formula

$$a(\hat{\phi}_k) = \frac{1 - (P_n g^{-1}(X_{A_k}^T \hat{\theta}_k)^2 - \{P_n g^{-1}(X_{A_k}^T \hat{\theta}_k)\}^2)}{P_n b''\{h(X_{A_k}^T \hat{\theta}_k)\}},$$

where $h(\cdot) = b'^{-1}(g^{-1}(\cdot))$ and P_n is the empirical measure with the reference data. For normal family where the canonical link is an identity function, we have $b''(\psi) = 1$, which implies the denominator is 1.

A.3 Full proof of theorem 2.2.1 and Checking regularity assumptions in two examples

A.3.1 Regulartity Assumptions for Theorem 2.2.1

Assumptions (A1)-(A4) are for consistency and the additional assumptions (A5)-(A9) are for asymptotic normality.

(A1): C is positive semi-definite and $CE\{U(X;\beta,\theta^*)\} = 0$ if and only if $\beta = \beta^*$.

(A2): $\beta^* \in D_\beta$, which is compact.

(A3): $u_k(X;\beta,\theta_k)$ is continuous for each $(\beta,\theta_k) \in D_\beta \times \mathcal{N}(\theta_k^*)$ with probability one, where $\mathcal{N}(\theta_k^*)$ is a neighborhood of θ_k^* for $k = 1, \ldots, K$.

(A4): $E\{\sup_{(\beta,\theta_k)\in D_\beta \times \mathcal{N}(\theta_k^*)} ||u_k(X;\beta,\theta_k)||\} < \infty$ for $k = 1, \ldots, K$.

(A5): $\partial u_k(X;\beta,\theta_k)/\partial\beta$ is continuous at each $(\beta,\theta_k) \in \mathcal{N}(\beta^*) \times \mathcal{N}(\theta_k^*)$ with probability 1, where $N(\beta^*)$ is a neighborhood of β^*.

(A6): $E\{\sup_{(\beta,\theta_k)\in \mathcal{N}(\beta^*) \times \mathcal{N}(\theta_k^*)} ||\partial u_k(X,\beta,\theta_k)/\partial\beta||\} < \infty$.

(A7): $\partial u_k(X;\beta^*,\theta_k)/\partial\theta_k$ is continuous at each $\theta_k \in \mathcal{N}(\theta_k^*)$ with probability one.

(A8): $E\{\sup_{\theta_k \in \mathcal{N}(\theta_k^*)} ||\partial u_k(X, \beta^*, \theta_k)||/\partial \theta_k\} < \infty$.

(A9): $\Delta(\beta^*, \theta^*)$ exists and is finite and $\Gamma(\beta^*, \theta^*)$ is of full rank.

More on the global identification assumption (A1) : Sometimes it's difficult to practically check the global identification condition. This motivates us to investigate conditions for local identifiability, or equivalently, the invertibility of the matrix of second derivatives at the true parameter, i.e., $\partial^2 Q(\beta)/\partial \beta^2 \mid_{\beta=\beta^*} = [E\{\partial U(X;\beta)/\partial \beta\}^T C E\{\partial U(X;\beta)/\partial \beta\}] \mid_{\beta=\beta^*}$ [142, 51], assuming C is a positive definite matrix. The condition can be stated in terms of the equivalent sample version of the matrix, given by, $X_{rbind}^T W X_{A_{diag}} C X_{A_{diag}}^T W X_{rbind}$. As C is a positive definite matrix, the entire local identifiability condition for the sample version then boils down to $X_{A_{diag}}^T W X_{rbind}$ being a full column rank matrix. A sufficient condition for this is $X_{A_{diag}}$ contains information on all the covariates of the maximal model. In other words, the individual covariates in the maximal model have to be part of at least one of the reduced models.

We first provide a complete proof of Theorem 1 and then check the assumptions for logistic and linear regression models.

Proof of Theorem 2.2.1 : First, we show the consistency of $\hat{\beta}$. Denote $\hat{\theta}$ and θ^* as stacked vectors of $\hat{\theta}_k$'s and θ_k^*'s, respectively. Denote $U_0(\beta, \theta) = E(U(X;\beta,\theta))$ and $Q_0(\beta) = U_0(\beta, \theta^*)^T C U_0(\beta, \theta^*)$.

By (A1) and Lemma 2.3 of [124], $Q_0(\beta)$ is uniquely minimized at β^*.

By (A2), (A3), (A4) and Lemma 2.4 of [124], $U_0(\beta, \theta)$ is continuous and $U_n(\beta, \theta)$ converges uniformly to $U_0(\beta, \theta)$ for $(\beta, \theta) \in D_\beta \times N_c(\theta^*)$, where $N_c(\theta^*)$ is a compact subset of $N(\theta^*)$ including θ^*. Note that $\hat{\theta}$ is a consistent estimator of θ^*. With probability going to one (wpg1),

$$\sup_{\beta \in D_\beta} ||U_n(\beta, \hat{\theta}) - U_0(\beta, \hat{\theta})|| \leq \sup_{(\beta,\theta) \in D_\beta \times N_c(\theta^*)} ||U_n(\beta, \theta) - U_0(\beta, \theta)||.$$

Then, $U_n(\beta, \hat{\theta}) - U_0(\beta, \hat{\theta})$ converges uniformly in probability to 0 for $\beta \in D_\beta$.

For any $r > 0$, wpg1,

$$\sup_{\beta \in D_\beta} ||U_0(\beta, \hat{\theta}) - U_0(\beta, \theta^*)|| \leq \sup_{\beta \in D_\beta} E(\sup_{||\theta - \theta^*|| < r} ||U(\beta, \theta) - U(\beta, \theta^*)||).$$

By (A3), (A4) and dominant convergence theorem, $E(\sup_{||\theta-\theta^*||<r} ||U(\beta, \theta) - U(\beta, \theta^*)||)$ converges to 0 for every $\beta \in D_\beta$ as r decreases to 0. Note that $E(\sup_{||\theta-\theta^*||<r} ||U(\beta, \theta) - U(\beta, \theta^*)||)$ decreases as r decreases for each β. By (A2) and Dini's theorem (see, for example, Theorem 7.13 of [143]), $E(\sup_{||\theta-\theta^*||<r} ||U(\beta, \theta) - U(\beta, \theta^*)||)$ converges uniformly in probability to 0 for $\beta \in D_\beta$ as r decreases to 0. Then, $U_0(\beta, \hat{\theta}) - U_0(\beta, \theta^*)$ converges uniformly in probability to 0 for $\beta \in D_\beta$.

By combining the above two results, it follows that $U_n(\beta, \hat{\theta})$ converges uniformly in probability to $U_0(\beta, \theta^*)$ for $\beta \in D_\beta$.

By the triangle and Cauchy-Schwartz inequalities,

$$\begin{aligned}\sup_{\beta \in D_\beta} |Q_n(\beta) - Q_0(\beta)| \leq{}& ||\hat{C}|| \sup_{\beta \in D_\beta} ||U_n(\beta, \hat{\theta}) - U_0(\beta, \theta^*)||^2 \\ &+ 2||\hat{C}|| \sup_{\beta \in D_\beta} ||U_0(\beta, \theta^*)|| \sup_{\beta \in D_\beta} ||U_n(\beta, \hat{\theta}) - U_0(\beta, \theta^*)|| \\ &+ ||\hat{C} - C|| \sup_{\beta \in D_\beta} ||U_0(\beta, \theta^*)||^2\end{aligned}$$

Since $\hat{C}$ is a consistent estimator of C, $||\hat{C}||$ converges in probability to $||C||$, which is finite; $||\hat{C} - C||$ converges in probability to 0. Since $U_0(\beta, \theta^*)$ is continuous for $\beta \in D_\beta$ and D_β is compact, $\sup_{\beta \in D_\beta} ||U_0(\beta, \theta^*)||^2$ is finite. Since $\sup_{\beta \in D_\beta} ||U_n(\beta, \hat{\theta}) - U_0(\beta, \theta^*)||$ converges in probability to 0, $\sup_{\beta \in D_\beta} ||U_n(\beta, \hat{\theta}) - U_0(\beta, \theta^*)||^2$ converges in probability to 0. Thus, $Q_n(\beta) - Q_0(\beta)$ converges uniformly in probability to 0 for $\beta \in D_\beta$. Recall that β^* is the unique minimizer of $Q_0(\beta)$. By Theorem 2.1 of [124], $\hat{\beta}$ is a consistent estimator of β^*.

Next, we derive the asymptotic distribution of the GENMETA estimator $\hat{\beta}$. Note that $\hat{\beta}$ is a solution to

$$G_n(\beta, \hat{\theta})^T \hat{C} U_n(\beta, \hat{\theta}) = 0,$$

where $G_n(\beta, \hat{\theta}) = \partial U_n(\beta, \hat{\theta})/\partial \beta$, the Jacobian of $U_n(\beta, \hat{\theta})$. On the other hand, by mean value theorem,

$$U_n(\hat{\beta}, \hat{\theta}) = U_n(\beta^*, \hat{\theta}) + G_n(\bar{\beta}, \hat{\theta})(\hat{\beta} - \beta^*),$$

where $\bar{\beta}$ denotes a matrix each column of which corresponds to each element of $U_n(\beta, \hat{\theta})$. After left multiplying $G_n(\hat{\beta}, \hat{\theta})^T \hat{C}$ to the above identity, it follows

$$n^{1/2}(\hat{\beta} - \beta^*) = -M_n n^{1/2} U_n(\beta^*, \hat{\theta}),$$

where $M_n = (G_n(\hat{\beta}, \hat{\theta})^T \hat{C} G_n(\bar{\beta}, \hat{\theta}))^{-1} G_n(\hat{\beta}, \hat{\theta})^T \hat{C}$.

Consider M_n. Since $\hat{\beta}$ is a consistent estimator of β^*, each column of $\bar{\beta}$ is a consistent estimator of β^*. On the other hand, $\hat{\theta}$ is a consistent estimator of θ^*. By (A5), (A6) and Lemma 2.4 of [124], $G_n(\beta, \theta)$ converge uniformly to continuous $E\{\partial U(X;\beta,\theta)/\partial\beta\}$ for $(\beta,\theta) \in D_\beta \times N_c(\theta^*)$, where $N_c(\theta^*)$ is a compact subset of $N(\theta^*)$, including θ^*. Since $\hat{\beta}$ and each column of $\bar{\beta}$ converge in probability to β^* and $\hat{\theta}$ is a consistent estimator of θ^*, by, for example, Theorem 9.4 of [89], both $G_n(\hat{\beta},\hat{\theta})$ and $G_n(\bar{\beta},\hat{\theta})$ converges in probability to $\Gamma = E\{\partial U(X;\beta^*,\theta^*)/\partial\beta\}$. Thus, by noting $\hat{C}{\rightarrow}C$ in probability, M_n converges in probability to $(\Gamma^T C\Gamma)^{-1}\Gamma^T C$.

Consider $n^{1/2}U_n(\beta^*,\hat{\theta})$. By mean value theorem,

$$U_n(\beta^*,\hat{\theta}) = U_n(\beta^*,\theta^*) + V_n(\beta^*,\bar{\theta})(\hat{\theta}-\theta^*),$$

where V_n is the Jacobian of $U_n(\beta^*,\theta)$ as a function of θ and $\bar{\theta}$ is a matrix each column of which corresponds to each element of $U_n(\beta^*,\theta)$. Thus,

$$n^{1/2}U_n(\beta^*,\hat{\theta}) = n^{1/2}U_n(\beta^*,\theta^*) + V_n(\beta^*,\bar{\theta})n^{1/2}(\hat{\theta}-\theta^*).$$

By (A9) and central limit theorem, $n^{1/2}U_n(\beta^*,\theta^*) \overset{d}{\rightarrow} N(0,\Delta)$. Since $\hat{\theta}$ is a consistent estimator of θ^*. each column of $\bar{\theta}$ converges in probability to θ^*. Similar to the above argument, by (A7), (A8), Lemma 2.4 of [124] and Theorem 9.4 of [89],

$$V_n(\beta^*,\bar{\theta}){\rightarrow}\mathrm{diag}(W_1,W_2,\ldots,W_K) \quad \text{in probability,}$$

where, for $k=1,2,\ldots,K$, $W_k = E\{\partial u_k(X,\beta^*,\theta_k)/\partial\theta_k\} \mid_{\theta_k=\theta_k^*}$. The K study data sets are independent. So are $\hat{\theta}_k$'s. Note that $n_k/n \rightarrow c_k$, where c_k is a positive constant for $k=1,2,\ldots,K$. Then $n^{1/2}(\hat{\theta}-\theta^*)$ converges in distribution to

$$N(0, \mathrm{diag}((1/c_1)\Sigma_1, (1/c_2)\Sigma_2, \ldots, (1/c_K)\Sigma_K)).$$

Since the K data sets and the reference data are independent, the above results imply that $n^{1/2}U_n(\beta^*, \hat{\theta})$ converges in distribution to $N(0, \Delta + \Lambda)$, where Λ is a block diagonal matrix whose kth block is $(1/c_k)W_k\Sigma_k W_k^T$ for $k = 1, \ldots, K$.

Therefore, with the above two results on M_n and $n^{1/2}U_n(\beta^*, \hat{\theta})$ and by Slutsky's theorem, the asymptotic normality of $n^{1/2}(\hat{\beta} - \beta^*)$ follows.

Checking assumptions for logistic regression model : Suppose the maximal model is

$$Y \mid X \sim \text{Bernoulli}\Big\{\frac{1}{1 + \exp(-X^T\beta^*)}\Big\},$$

where $X = (1, X^T)^T$, $X = (X_1, \ldots, X_d)^T$ is the vector of covariates and $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^T$ is the vector of coefficients of interest. There are K independent studies and the reduced model of the kth study is

$$Y \mid X_{A_k} \sim \text{Bernoulli}\Big\{\frac{1}{1 + \exp(-X_{A_k}^T\theta_k)}\Big\},$$

where $X_{A_k} = (1, X_{A_k}^T)^T$, X_{A_k} is a sub-vector of X with $A \subset \{1, 2, \ldots, p\}$. For example, $X_A = (X_1, X_2)^T$ when $A = \{1, 2\}$.

The global identification assumption (A1) usually holds and D_β is a compact set. Next, we check the assumptions (A3) to (A9). The moment functions from the kth study is

$$u_k(X; \beta, \theta_k) = \Big(\frac{1}{1 + e^{-X^T\beta}} - \frac{1}{1 + e^{-X_{A_k}^T\theta_k}}\Big)X_{A_k}.$$

It is a continuous function of β and θ_k. Then, (A3) is satisfied. Note that

$$\sup_{(\beta,\theta)\in D_\beta\times N(\theta^*)} ||\Big(\frac{1}{1+e^{-X^T\beta}} - \frac{1}{1+e^{-X_{A_k}^T\theta_k}}\Big)X_{A_k}|| \leq 2||X||_1,$$

where $||\cdot||$ and $||\cdot||_1$ are the l_2 and l_1 norms, respectively. Then, given $E(|X_i|) < \infty$ for each i, (A4) is satisfied. Also,

$$\frac{\partial}{\partial\beta}u_k(X;\beta,\theta_k) = \frac{e^{-X^T\beta}}{(1+e^{-X^T\beta})^2}X_{A_k}X^T, \tag{A.13}$$

which does not depend on θ_k and is continuous for each β. Then, (A5) is verified. Note that

$$\sup_{(\beta,\theta)\in D_\beta\times N(\theta^*)} ||\frac{e^{-X^T\beta}}{(1+e^{-X^T\beta})^2}X_{A_k}X^T|| \leq ||XX^T||_1.$$

Given $E(X_i^2) < \infty$ for each i, (A6) is satisfied. Note that

$$\frac{\partial}{\partial\theta_k}u_k(X;\beta^*,\theta_k) = -\frac{e^{-X_{A_k}^T\theta_k}}{(1+e^{-X_{A_k}^T\theta_k})^2}X_{A_k}X_{A_k}^T,$$

which is continuous for each θ_k. Then, (A7) is satisfied. Note that

$$\sup_{(\beta,\theta)\in D_\beta\times N(\theta^*)} ||-\frac{e^{-X_{A_k}^T\theta_k}}{(1+e^{-X_{A_k}^T\theta_k})^2}X_{A_k}X_{A_k}^T|| \leq ||XX^T||_1.$$

Given $E(X_i^2) < \infty$ for each i, (A8) is satisfied. The absolute value of each element of $\Delta(\beta^*,\theta^*)$ is less than 1, $E(|X_i|)$ or $E(|X_iX_j|)$ for each i and j. Given $E(X_i^2) < \infty$, $\Delta(\beta^*,\theta^*)$ is finite. Note that $\Gamma(\beta^*,\theta_k^*)$ is a stacked matrix of (A.13) for $k = 1,\dots,K$. Given each covariate of the maximal model is in at least one reduced model and $E[\{e^{-X^T\beta}/(1+e^{-X^T\beta})^2\}XX^T]$ is positive definite, $\Gamma(\beta^*,\theta^*)$

is of full rank. Then, (A9) is verified. □

Checking assumptions for linear regression Model : Suppose the true maximal model is

$$Y \mid X \sim N(X^T\beta^*, \sigma^{*2}),$$

where $X = (X_1, X_2, \dots, X_p)^T$; $\beta^* = (\beta_1^*, \beta_2^*, \dots, \beta_p^*)^T$; $E(X) = 0$ and $E(Y) = 0$, that is, both X and Y are centered. There are K independent studies and the reduced model of the kth study is

$$Y \mid X_{A_k} \sim N(X_{A_k}^T\theta_k, \sigma_k^2).$$

For simplicity, assume σ^{*2} is known and the unknown parameter is β^*. The case with unknown σ^{*2} can be similarly considered.

The moment functions from the kth reduced model is

$$u_k(X;\beta;\theta_k,\sigma_k^2) = \frac{1}{\sigma_k^2}(X_{A_k}X^T\beta - X_{A_k}X_{A_k}^T\theta_k),$$

which is linear in β. Note that

$$\frac{\partial}{\partial\beta}u_k(X;\beta;\theta_k,\sigma_k^2) = \frac{1}{\sigma_k^2}X_{A_k}X^T. \tag{A.14}$$

Given each covariate of the maximal model is in at least one reduced model and $E(XX^T)$ is positive definite, $\Gamma(\beta^*, \{\theta_k^*\}, \{\sigma_k^{*2}\}) = \partial u_k(X;\beta^*;\{\theta_k^*\},\{\sigma_k^{*2}\})/\partial\beta$ is of full rank. Given C is positive definite, (A1) is satisfied. Suppose D_β is a compact set. Then, (A2) is satisfied.

Next, we check the assumptions (A3) to (A9). Note that $u_k(X;\beta;\theta_k,\sigma_k^2)$ is continuous for every $(\beta, \theta_k, \sigma_k^2)$. Then, (A3) is satisfied. Note that

$$\sup_{(\beta,\theta_k,\sigma_k^2)} ||\frac{1}{\sigma_k^2}(X_{A_k}X^T\beta - X_{A_k}X_{A_k}^T\theta_k)|| \leq \frac{1}{\sigma_k^2}(||\beta|| + ||\theta_k||)||XX^T||_1,$$

Denote a finite upper bound of $||\beta||$ for $\beta \in D_\beta$ as $C(\beta)$, a finite upper bound of $||\theta_k||$ for $\theta_k \in N(\theta_k^*)$ as $C(\theta_k)$, and a positive finite lower bound of σ_k^2 for $\sigma_k^2 \in N(\theta_k^*)$ as σ_L^2. The supremum of $(1/\sigma_k^2)(||\beta|| + ||\theta_k||)$ for $(\beta, \theta_k, \sigma_k^2) \in D_\beta \times N(\theta_k^*) \times N(\sigma_k^{*2})$ is bounded by $(1/\sigma_L^2)(C(\beta) + C(\theta_k))$. Given $E(X_i^2) < \infty$ for each i, (A4) is satisfied. Note that $\partial u_k(X; \beta; \theta_k, \sigma_k^2)/\partial\beta$ does not depend on β and θ_k and is continuous for each σ_k^2. Then, (A5) is satisfied. Note that

$$\sup_{\sigma_k^2 \in N(\sigma_k^{*2})} ||\frac{1}{\sigma_k^2}X_{A_k}X^T|| \leq \frac{1}{\sigma_L^2}||XX^T||_1.$$

Given $E(X_i^2) < \infty$ for each i, (A6) is satisfied. Note that

$$\frac{\partial}{\partial(\theta_k, \sigma_k^2)}u_k(X; \beta; \theta_k, \sigma_k^2) = \{-\frac{1}{\sigma_k^2}X_{A_k}X_{A_k}^T, -\frac{1}{\sigma_k^4}(X_{A_k}X^T\beta - X_{A_k}X_{A_k}^T\theta_k)\},$$

which is continuous for every $(\beta, \theta_k, \sigma_k^2)$. Then, (A7) is satisfied. For every $(\beta, \theta_k, \sigma_k^2) \in D_\beta \times N(\theta_k^*, N(\sigma_k^{*2}))$, the l_2 norm of the above partial derivative is less than or equal to

$$\frac{1}{\sigma_L^2} + \frac{1}{\sigma_L^4}(C(\beta) + C(\theta_k))||XX^T||_1.$$

Given $E(X_i^2) < \infty$ for each i, (A8) is satisfied. Each element of $\Delta(\beta^*, \{\theta_k^*\}, \{\sigma_k^{*2}\})$ is equal to a constant times $E(X_{i_1}X_{i_2}X_{i_3}X_{i_4})$ for some i_1, i_2, i_3, i_4. Given $E(X_i^4) < \infty$ for each i, Δ is finite. Note that $\Gamma(\beta^*, \{\theta_k^*\}, \{\sigma_k^{*2}\})$ is a stacked matrix of

(A.14) for $k = 1, \ldots, K$. As in checking (A2), given each covariate of the maximal model is in at least one reduced model and $E(XX^T)$ is positive definite, Γ is of full rank. Then, (A9) is verified. $\square$

A.4 Simulation Results for Log-normally Distributed Covariates

Table A.1: Robustness of GENMETA Estimation (Log-normally Distributed Covariates)

Setting	Study-I	Study-II	Study-III	Reference	β_i^*	Bias	SD (ESD)	RMSE	CR	AL
I	μ_b	μ_b	μ_b	μ_b	β_1^*	.010	.076 (.075)	.077	.941	.288
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.011	.064 (.061)	.065	.947	.237
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	.006	.066 (.064)	.066	.954	.246
II	μ_b	μ_h	μ_m	μ_b	β_1^*	.010	.079 (.072)	.079	.930	.272
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.002	.056 (.054)	.056	.948	.211
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	-.002	.062 (.058)	.062	.945	.222
	μ_b	μ_b	μ_b	μ_b	β_1^*	.032	.088 (.088)	.094	.930	.339
	σ_b^2	σ_h^2	σ_l^2	σ_b^2	β_2^*	-.002	.062 (.057)	.062	.941	.221
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	-.005	.074 (.074)	.074	.967	.286
	μ_b	μ_h	μ_m	μ_b	β_1^*	.021	.079 (.077)	.081	.929	.294
	σ_b^2	σ_h^2	σ_l^2	σ_b^2	β_2^*	.0005	.055 (.055)	.055	.956	.213
	ρ_b	ρ_b	ρ_b	ρ_b	β_3^*	-.008	.065 (.064)	.065	.954	.246
III	μ_b	μ_b	μ_b	μ_b	β_1^*	-.062	.107 (.118)	.124	.934	.382
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.021	.070 (.065)	.073	.930	.250
	ρ_b	ρ_b	ρ_b	ρ_h	β_3^*	.030	.087 (.096)	.092	.956	.322
	μ_b	μ_b	μ_b	μ_b	β_1^*	.039	.072 (.069)	.081	.891	.264
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.023	.065 (.062)	.069	.932	.240
	ρ_b	ρ_b	ρ_b	ρ_l	β_3^*	.018	.061 (.058)	.064	.930	.224
	μ_b	μ_b	μ_b	μ_b	β_1^*	.053	.079 (.075)	.095	.866	.290
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.019	.065 (.063)	.067	.942	.242
	ρ_l	ρ_b	ρ_h	ρ_l	β_3^*	.012	.068 (.064)	.069	.935	.249
	μ_b	μ_b	μ_b	μ_b	β_1^*	.032	.089 (.084)	.095	.912	.322
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.010	.062 (.062)	.063	.946	.240
	ρ_l	ρ_b	ρ_h	ρ_b	β_3^*	-.009	.073 (.071)	.073	.942	.273
	μ_b	μ_b	μ_b	μ_b	β_1^*	-.025	.113 (.108)	.116	.954	.407
	σ_b^2	σ_b^2	σ_b^2	σ_b^2	β_2^*	.017	.065 (.064)	.067	.951	.248
	ρ_l	ρ_b	ρ_h	ρ_h	β_3^*	-.002	.091 (.091)	.091	.965	.347
IV			μ_b	μ_b	β_1^*	.007	.096 (.104)	.096	.968	.365
	$X_1 > -0.5$,	$X_2 > 0$	σ_b^2	σ_b^2	β_2^*	.242	.353 (.117)	.428	.572	.401
	$X_2 < 0.5$		ρ_b	ρ_b	β_3^*	-.015	.067 (.081)	.068	.971	.283

Biases, standard deviation (SD), estimated standard deviation (ESD), square roots of mean square errors (RMSE), coverage rates (CR), and average lengths (AL) of 95% confidence intervals of the GENMETA estimates using the study covariance estimators in the setting of logistic regression. In setting (I), data are simulated in ideal setting where the covariate distribution is a log-normal distribution with the natural logarithm of the covariates being characterized by mean, sd and correlation of normal variates and are assumed to same across all populations. In setting (II)-(IV), the assumption is violated by creating variations in mean/sd, correlations of the underlying normal distribution and selection criterion across the studies and reference sample. The vector of means, variances and correlations of the underlying normal covariates are denoted by $\mu_* = (\mu_1, \mu_2, \mu_3)$, $\sigma_*^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ and $\rho_* = (\rho_{12}, \rho_{23}, \rho_{13})$ for $* \in \{b, l, m, h\}$, where $\mu_b = (0, 0, 0)$, $\mu_m = (0.5, 0.5, 0.5)$, $\mu_h = (1, 1, 1)$; $\sigma_b^2 = (1, 1, 1)$, $\sigma_l^2 = (0.5, 0.5, 0.5)$, $\sigma_h^2 = (2, 2, 2)$ and $\rho_b = (0.3, 0.6, 0.1)$, $\rho_h = (0.4, 0.8, 0.2)$, $\rho_l = (0.2, 0.4, 0)$. Estimated standard deviation are obtained by the asymptotic formula (2) in the main paper and used to construct 95% confidence interval.

Appendix B

Chapter 2

In this section, we denote the vectors/matrices by bold symbols. We replace $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^{(R)}$ to denote the reduced parameter in phase-I.

B.1 Estimating equation in phase-I

Let $\boldsymbol{S}_{\boldsymbol{\theta}^{(R)}}(Y, \boldsymbol{X}^{(I)}) = (Y - expit(\boldsymbol{\theta}^{(R)^T}\boldsymbol{X}^{(I)}))\boldsymbol{X}^{(I)}$ denote the score vector of dimension q_1. Also, let $\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)}) := expit(\boldsymbol{\beta}^T\boldsymbol{X}_i^{(II)}) - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{X}_i^{(I)})\boldsymbol{X}_i^{(I)}$ .Then,

$$E^{(I)}(\boldsymbol{S}_{\boldsymbol{\theta}^{(R)}}(Y, \boldsymbol{X}^{(I)})) = 0$$

$$\begin{aligned}
E^{(I)}(\boldsymbol{S}_{\boldsymbol{\theta}^{(R)}}(Y, \boldsymbol{X}^{(I)})) &= E^{(I)}_{\boldsymbol{X},S}[E_{Y|\boldsymbol{X},S}\{(Y - expit(\boldsymbol{\theta}^{(R)^T}\boldsymbol{X}^{(I)}))\boldsymbol{X}^{(I)}\}] \\
&= \sum_{s=1}^{J}\sum_{i=1}^{n}[\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})Pr(\boldsymbol{X}_i = \boldsymbol{x}_i, S_i = s_i)] \\
&= \sum_{d=0}^{1}\sum_{s=1}^{J}\sum_{i=1}^{n}[\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})Pr(\boldsymbol{X}_i = \boldsymbol{x}_i, S_i = s, Y_i = d)] \\
&= \sum_{d=0}^{1}\sum_{s=1}^{J}\sum_{i=1}^{n}[\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})Pr(\boldsymbol{X}_i = \boldsymbol{x}_i | S_i = s, Y_i = d)Pr(S_i = s, Y_i = d)] \\
&= \sum_{d=0}^{1}\sum_{s=1}^{J}\sum_{i=1}^{n}[\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})\frac{1}{n_{ds}}\frac{N_{ds}}{N}\mathbf{1}_{(y_i=d, s_i=s)}] \\
&= \frac{1}{N}\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{s=1}^{J}[\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})\frac{N_{ds}}{n_{ds}}\mathbf{1}_{(y_i=d, s_i=s)}] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{d=0}^{1}\sum_{s=1}^{J}[R_i\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})\frac{N_{ds}}{n_{ds}}\mathbf{1}_{(y_i=d, s_i=s)}]
\end{aligned}$$

Let us denote $\sum_{d,s}\frac{N_{ds}}{n_{ds}}\mathbb{1}_{\{W_i \in \mathcal{W}_{ds}\}}$ by $\frac{1}{\pi(\boldsymbol{W}_i)}$ where $\pi(\boldsymbol{W}_i) = \sum_{d,s}\frac{n_{ds}}{N_{ds}}\mathbb{1}_{\{\boldsymbol{W}_i \in \mathcal{W}_{ds}\}}$. Then the above estimating equation can be rewritten as

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}^{(R)})}{\pi(\boldsymbol{W}_i)} = 0 \qquad \text{(B.1)}$$

B.2 Estimating equation in phase-II

Let $\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y, \boldsymbol{X}) = (Y - expit(\boldsymbol{\theta}^{(CC)^T}\boldsymbol{X}))\boldsymbol{X}$ denote the score vector of dimension q_2 where $\boldsymbol{\theta}^{(CC)} = \boldsymbol{\theta}^{(CC)}(\boldsymbol{\beta}) = \boldsymbol{\beta} + (\log \frac{p(1,s)}{p(0,s)}, \boldsymbol{0}^T)^T$Then,

$$E^{(II)}(\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y, \boldsymbol{X})) = 0 \tag{B.2}$$

$$\begin{aligned}
E^{(II)}(\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y, \boldsymbol{X})) &= E_Y[E^{(II)}_{\boldsymbol{X}|Y}\{\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y, \boldsymbol{X})\}] \\
&= \frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y_i = 1, \boldsymbol{X}_i) + \frac{n_0}{n}\frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y_i = 0, \boldsymbol{X}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{S}_{\boldsymbol{\theta}^{(CC)}}(Y_i, \boldsymbol{X}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - expit(\boldsymbol{\theta}^{(CC)^T}\boldsymbol{X}_i))\boldsymbol{X}_i \\
&= \frac{1}{n}\sum_{i=1}^{N}R_i(Y_i - expit(\boldsymbol{\theta}^{(CC)^T}\boldsymbol{X}_i^{(II)}))\boldsymbol{X}_i^{(II)} \\
&= \frac{1}{N}\frac{N}{n}\sum_{i=1}^{N}R_i(Y_i - expit(\boldsymbol{\theta}^{(CC)^T}\boldsymbol{X}_i^{(II)}))\boldsymbol{X}_i^{(II)}
\end{aligned}$$

B.3 Consistency and Asymptotic Normality

Consistency can be proved in a similar way as shown in our original GMeta paper. Asymptotic normality will also follow in the same direction, but, here, we need to take into account of the dependence between the two phases.

Let $\boldsymbol{G}_N(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}_N(\boldsymbol{\beta})$. Then, $\boldsymbol{G}_N^T(\boldsymbol{\beta}) C \boldsymbol{U}_N(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{GMeta}} = 0$. From now on, we will denote $\hat{\boldsymbol{\beta}}_{GMeta}$ by $\hat{\boldsymbol{\beta}}$. By Mean-value theorem,

$$\sqrt{N}\boldsymbol{U}_N(\hat{\boldsymbol{\beta}}) = \sqrt{N}\boldsymbol{U}_N(\boldsymbol{\beta}_0) + \sqrt{N}\boldsymbol{G}_N(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

where $\bar{\boldsymbol{\beta}} \in (\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}})$. Pre-multiplying the above by $\boldsymbol{G}_N^T(\hat{\boldsymbol{\beta}})\boldsymbol{C}$, we get

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\boldsymbol{M}_N(\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(R)})\sqrt{N}\boldsymbol{U}_N(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}^{(R)}))$$

where $\boldsymbol{M}_N = \{\boldsymbol{G}_N^T(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(R)})\boldsymbol{C}\boldsymbol{G}_N(\bar{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(R)})\}^{-1}\boldsymbol{G}_N^T(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(R)})\boldsymbol{C}$. Assuming $\hat{\boldsymbol{\theta}}^{(R)}$ a consistent estimator for $\boldsymbol{\theta}_0$, we have, under some regularity conditions, $\boldsymbol{G}_N^T(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(R)}) \xrightarrow{P} \boldsymbol{\Gamma}^T$ and $\boldsymbol{G}_n(\bar{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(R)}) \xrightarrow{P} \boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} = E_{\boldsymbol{V},\boldsymbol{X}} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\theta}^{(R)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \boldsymbol{\theta}^{(R)}=\boldsymbol{\theta}_0}$. Then, $\boldsymbol{M}_N \xrightarrow{P} (\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{C}$. Focussing on the second multiplicative term, by mean value theorem, we have

$$\sqrt{N}\boldsymbol{U}_N(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}^{(R)}) = \sqrt{N}\boldsymbol{U}_N(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) + \sqrt{N}\boldsymbol{V}_N(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(R)} - \boldsymbol{\theta}_0)$$

where $\boldsymbol{V}_N(\bar{\boldsymbol{\theta}}) = (\boldsymbol{V}_{1N}^T(\boldsymbol{\beta}_0, \bar{\boldsymbol{\theta}}^{(R)}), \boldsymbol{0}^T)^T$, $\bar{\boldsymbol{\theta}}^{(R)} \in (\hat{\boldsymbol{\theta}}^{(R)},\ \boldsymbol{\theta}_0)$, $\boldsymbol{V}_{1N}(\boldsymbol{\beta}_0, \bar{\boldsymbol{\theta}}^{(R)}) = \frac{\partial}{\partial \boldsymbol{\theta}^{(R)}} \boldsymbol{U}_{1N}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^{(R)})|_{\boldsymbol{\theta}^{(R)}=\bar{\boldsymbol{\theta}}^{(R)}}$.

Focussing on the phase-I moment vector of the first term of the above equation, we have

$$\sqrt{N}\boldsymbol{U}_{1N}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \sqrt{N}\boldsymbol{U}_{1N}^*(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) + \{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} R_i \boldsymbol{f}(\boldsymbol{X}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\} o_p(1)$$

where $\boldsymbol{U}_{1N}^*(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)}{p(\boldsymbol{W}_i)}$. Under regularity conditions $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} R_i \boldsymbol{f}(\boldsymbol{X}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$

$O_p(1)$. Then by Slutsky's theorem, we have

$$\sqrt{N}\boldsymbol{U}_{1N}(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) = \sqrt{N}\boldsymbol{U}^*_{1N}(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) + o_p(1)$$

By WLLN, we have $\pi(\boldsymbol{V}) = p(\boldsymbol{V}) + o_p(1)$. Using this we have $\boldsymbol{V}_{1N}(\boldsymbol{\beta}_0,\bar{\boldsymbol{\theta}}^{(R)}) = \boldsymbol{V}^*_{1N}(\boldsymbol{\beta}_0,\bar{\boldsymbol{\theta}}^{(R)}) + o_p(1)$ where $\boldsymbol{V}^*_{1N} = \frac{\partial \boldsymbol{U}^*_{1N}}{\partial \boldsymbol{\theta}^{(R)}}$. By WLLN, we have $\boldsymbol{V}^*_{1N}(\boldsymbol{\beta}_0,\bar{\boldsymbol{\theta}}^{(R)}) \xrightarrow{P} \boldsymbol{V}_1$ where $\boldsymbol{V}_1 = \boldsymbol{V}_1(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) = E\frac{\partial}{\partial\boldsymbol{\theta}^{(R)}}\boldsymbol{U}_1(\boldsymbol{\beta}_0,\boldsymbol{\theta}^{(R)})|_{\boldsymbol{\theta}^{(R)}=\boldsymbol{\theta}_0}$

Therefore, the influence function representation is given by,

$$\begin{aligned}
\sqrt{N}\boldsymbol{U}_N(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}}^{(R)}) &= \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{V}^*_{1N} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \sqrt{N} \begin{pmatrix} \boldsymbol{U}^*_{1N}(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) \\ \boldsymbol{U}_{2N}(\boldsymbol{\beta}_0) \\ \hat{\boldsymbol{\theta}}^{(R)} - \boldsymbol{\theta}_0 \end{pmatrix} + o_p(1) \\
&= \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{V}^*_{1N}\mathcal{I}_N^{-1}(\bar{\boldsymbol{\theta}}^{(R)}) \\ \boldsymbol{0} & \frac{N}{n}\boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}_1(\boldsymbol{W}_i,\boldsymbol{X}_i;\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}_2(\boldsymbol{W}_i,\boldsymbol{X}_i;\boldsymbol{\beta}_0) \\ \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}_3(Y_i,\boldsymbol{X}_i;\boldsymbol{\theta}_0) \end{pmatrix} + o_p(1) \\
&= \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{V}^*_{1N}\mathcal{I}_N^{-1}(\bar{\boldsymbol{\theta}}^{(R)}) \\ \boldsymbol{0} & \frac{N}{n}\boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{X}_i;\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) + o_p(1)
\end{aligned} \tag{B.3}$$

where,

$$\begin{gathered}
\boldsymbol{\Psi}_1(\boldsymbol{W}_i,\boldsymbol{X}_i;\boldsymbol{\beta}_0,\boldsymbol{\theta}_0) = \frac{R_i\boldsymbol{f}(\boldsymbol{X}_i,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)}{p(\boldsymbol{W}_i)} \\
\boldsymbol{\Psi}_2(\boldsymbol{W}_i,\boldsymbol{X}_i;\boldsymbol{\beta}_0) = R_i\boldsymbol{S}_{\boldsymbol{\beta}_0}(Y_i,\boldsymbol{X}_i^{(II)}) \\
\boldsymbol{\Psi}_3(Y_i,\boldsymbol{X}_i;\boldsymbol{\theta}_0) = \{Y_i - expit(\boldsymbol{\theta}_0^T\boldsymbol{X}_i^{(I)})\boldsymbol{X}_i^{(I)}\}
\end{gathered}$$

$$\mathcal{I}_N(\bar{\boldsymbol{\theta}}^{(R)}) \text{ is the information matrix}$$

By WLLN, $\mathcal{I}_N^{-1}(\bar{\boldsymbol{\theta}}^{(R)}) \xrightarrow{P} \mathcal{I}^{-1}(\boldsymbol{\theta}_0) = (E[(g^{-1})'(\boldsymbol{\theta}^{(R)^T}\boldsymbol{X}^{(I)})\boldsymbol{X}^{(I)}\boldsymbol{X}^{(I)^T}]|_{\boldsymbol{\theta}^{(R)}=\boldsymbol{\theta}_0})^{-1}$.

Let us assume $\frac{n}{N} \to \lambda \in (0,1)$. By central limit theorem, we have

$$\sqrt{N}\boldsymbol{U}_N(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}^{(R)}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Delta\Omega\Delta}^T)$$

where $\boldsymbol{\Omega} = E(\boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{X}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\boldsymbol{\Psi}^T(\boldsymbol{W}_i, \boldsymbol{X}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$, $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{I}_{q_1} & \mathbf{0} & \boldsymbol{V}_1\boldsymbol{\mathcal{I}}^{-1} \\ \mathbf{0} & \lambda^{-1}\boldsymbol{I}_{q_2} & \mathbf{0} \end{pmatrix}$.

Therefore, by Slutsky's theorem, we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, (\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{\Delta\Omega\Delta}^T\boldsymbol{C}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{\Gamma})^{-1})$$

Estimation of $\boldsymbol{\Omega}$, $\boldsymbol{V}_1$, $\boldsymbol{\mathcal{I}}$ from phase-II sample

$$\hat{\boldsymbol{\Omega}}_{11} = \textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{N_{ds}}{n_{ds}^2}\{expit(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j^{(II)}) - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{x}_j^{(I)})\}^2\boldsymbol{x}_j^{(I)}\boldsymbol{x}_j^{(I)^T}$$

$$\hat{\boldsymbol{\Omega}}_{22} = \textstyle\sum_{d,s} \frac{n_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}}\{y_j - expit(\hat{\boldsymbol{\theta}}^{(CC)^T}\boldsymbol{x}_j^{(II)})\}^2\boldsymbol{x}_j^{(II)}\boldsymbol{x}_j^{(II)^T}$$

$$\hat{\boldsymbol{\Omega}}_{33} = \textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}}\{y_j - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{x}_j^{(I)})\}^2\boldsymbol{x}_j^{(I)}\boldsymbol{x}_j^{(I)^T}$$

$$\hat{\boldsymbol{\Omega}}_{12} = \textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}}\{(expit(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j^{(II)}) - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{x}_j^{(I)}))(y_j - expit(\hat{\boldsymbol{\theta}}^{(CC)^T}\boldsymbol{x}_j^{(II)})\}\boldsymbol{x}_j^{(I)}\boldsymbol{x}_j^{(II)^T}$$

$$\hat{\boldsymbol{\Omega}}_{23} = \textstyle\sum_{d,s} \frac{n_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}}\{(y_j - expit(\hat{\boldsymbol{\theta}}^{(CC)^T}\boldsymbol{x}_j^{(II)}))(y_j - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{x}_j^{(I)})\}\boldsymbol{x}_j^{(II)}\boldsymbol{x}_j^{(I)^T}$$

$$\hat{\boldsymbol{\Omega}}_{13} = \textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}}\{(expit(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j^{(II)}) - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{x}_j^{(I)}))(y_j - expit(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{x}_j^{(I)})\}\boldsymbol{x}_j^{(I)}\boldsymbol{x}_j^{(I)^T}$$

$$\hat{\boldsymbol{V}}_1 = -\textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}} \frac{exp(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{X}_j^{(I)})}{(1+exp(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{X}_j^{(I)}))^2}\boldsymbol{X}_j^{(I)}\boldsymbol{X}_j^{(I)^T}$$

$$\hat{\boldsymbol{\mathcal{I}}} = \textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}} \frac{exp(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{X}_j^{(I)})}{(1+exp(\hat{\boldsymbol{\theta}}^{(R)^T}\boldsymbol{X}_j^{(I)}))^2}\boldsymbol{X}_j^{(I)}\boldsymbol{X}_j^{(I)^T}$$

$$\hat{\boldsymbol{\Gamma}}_1 = \textstyle\sum_{d,s} \frac{N_{ds}}{N} \sum_{j=1}^{n_{ds}} \frac{1}{n_{ds}} \frac{exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{X}_j^{(II)})}{(1+exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{X}_j^{(II)}))^2}\boldsymbol{X}_j^{(I)}\boldsymbol{X}_j^{(II)^T}$$

$$\hat{\mathbf{\Gamma}}_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{exp(\hat{\boldsymbol{\theta}}^{(CC)^T}\boldsymbol{X}_i^{(II)})}{(1+exp(\hat{\boldsymbol{\theta}}^{(CC)^T}\boldsymbol{X}_i^{(II)}))^2}\boldsymbol{X}_i^{(II)}\boldsymbol{X}_i^{(II)^T}$$

$$\mathbf{\Gamma} = (\mathbf{\Gamma}_1^T, \mathbf{\Gamma}_2^T)^T$$

Appendix C

Chapter 3

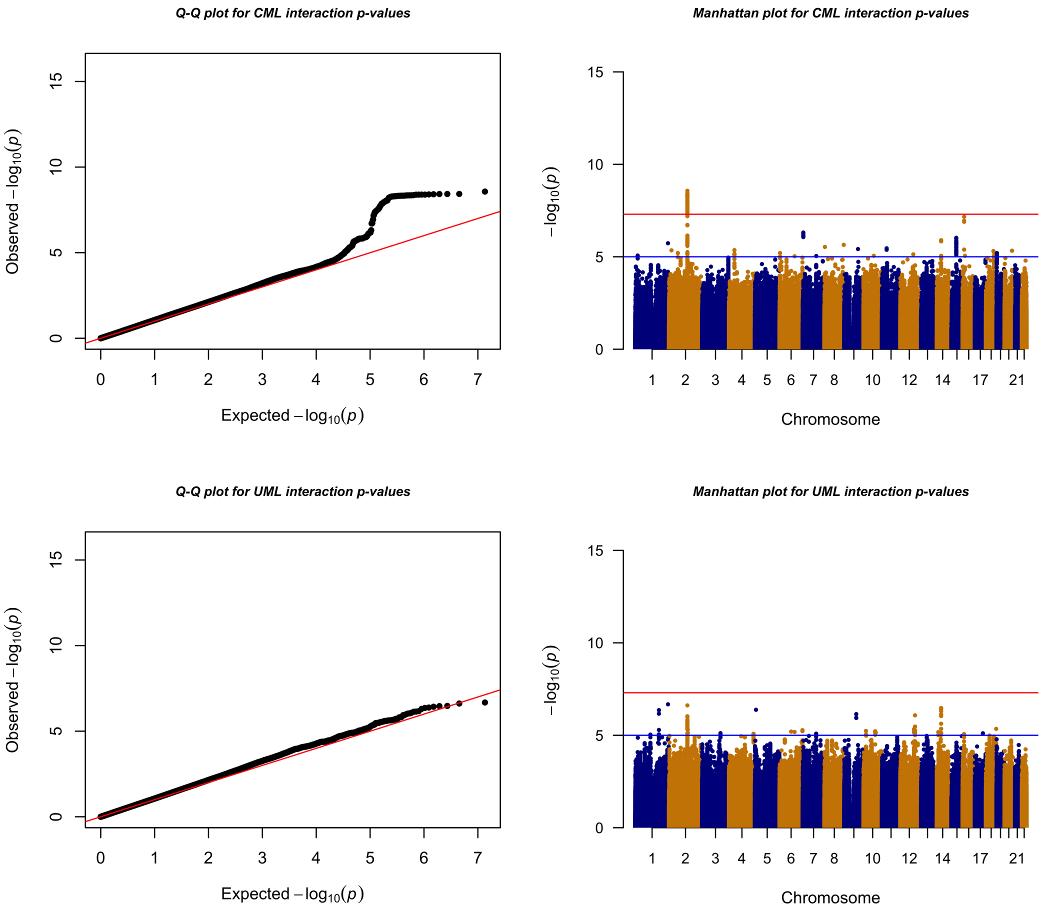


Figure C.1: Q-Q and Manhattan Plots of Interaction Analysis using CML and UML approach

Table C.1: Top-1 SNP (rs1818613) for locus 2q21.3 with association reaching genome wide significance

Chr BP SNP A1/A2 Gene	Dataset	Info	MAF	Analytical Method	OR SNP (95% CI)	OR SNP × FormerSmoker (95% CI)	OR SNP × CurrentSmoker (95% CI)	Interaction P-value
2q21.3 135356285 rs1818613 G/T TMEM163 (intronic)	Meta Analysis			CML	1.15 (1.1,1.21)	0.90 (0.83,0.97)	0.75 (0.66,0.84)	2.70E-09
				EB	1.15 (1.08,1.21)	0.88 (0.79,0.97)	0.74 (0.65,0.84)	3.08E-09
				UML	1.12 (1.05,1.18)	0.90 (0.8,0.99)	0.72 (0.59,0.84)	1.02E-06
	PanScan	0.99	0.39	CML	1.14 (1.06,1.21)	0.88 (0.79,0.97)	0.72 (0.6,0.84)	8.22E-07
				EB	1.11 (1.02,1.2)	0.90 (0.78,1.03)	0.72 (0.6,0.84)	1.01E-06
				UML	1.09 (1,1.18)	0.93 (0.81,1.06)	0.71 (0.55,0.87	1.24E-04
	PanC4	0.99	0.39	CML	1.18 (1.09,1.27)	0.92 (0.82,1.03)	0.79 (0.65,0.92)	2.49E-03
				EB	1.18 (1.09,1.27)	0.89 (0.73,1.04)	0.78 (0.63,0.93)	4.96E-03
				UML	1.17 (1.07,1.27)	0.85 (0.7,1)	0.76 (0.55,0.96)	1.30E-02

Abbreviations: Chr, chromosome; BP, base pair position according to the human genome Build 37; A1, effect allele (minor allele); A2, alternative allele (major allele) Info, imputation quality score; MAF, minor allele frequency; CML, Constrained maximum-likelihood; EB, Empirical Bayes; UML Unconstrained maximum-likelihood; OR (95%CI), odds ratios and confidence intervals.

Table C.2: Odds Ratios and Interaction P-value GWASs loci in Caucasian population.

Chr SNP Position^a Gene	Effect Allele/ Ref Allele	MAF PanC4 - PanScan	INFO^b PanC4 - PanScan	Method GxE	OR SNP (95% CI)	OR SNP × FormerSmokers (95% CI)	OR SNP × CurrentSmokers (95% CI)	Interaction P-value
1q32.1 rs2816938 199,985,368 NR5A2	A/T	.26-.24	.99-.99	CML	1.19 (1.12-1.27)	.99 (.91-1.08)	1.05 (.94-1.17)	0.58
				UML	1.18 (1.08-1.29)	.98 (.88-1.09)	1.16 (1.01-1.33)	0.06
				EB	1.19 (1.12-1.27)	.99 (.90-1.08)	1.09 (.95-1.26)	0.33
1q32.1 rs3790844 200,007,432 NR5A2	G/A	.22- .22	1-1	CML	.83 (.78-.88)	.94 (.87-1.03)	1.02 (.92-1.14)	0.23
				UML	.83 (.76-.91)	.96 (.86-1.08)	.99 (.85-1.14)	0.81
				EB	.83 (.78-.89)	.94 (.86-1.03)	1.01 (.90-1.13)	0.37
2p13.3 rs1486134 67,639,769 ETAA1 2236bp 3'	G/T	.289-.286	1-1	CML	.91 (.86-.97)	.99 (.93-1.06)	.99 (.91-1.08)	0.97
				UML	.89 (.84-.95)	1.01 (.90-1.12)	1.04 (.91-1.20)	0.8
				EB	.91 (.85-.96)	.99 (91-1.09)	1.00 (.90-1.12)	0.97
3q29 rs9854771 189,508,471 TP63	A/G	.344-.354	1-.998	CML	.89 (.84-.95)	1.00 (.93-1.08)	.99 (.90-1.09)	0.96
				UML	.90 (.84-.95)	.99 (.89-1.09)	1.03 (.91-1.17)	0.82
				EB	.89 (.84-.95)	1.00 (.92-1.09)	1.00 (.91-1.11)	0.99
5p15.33 rs2736098 1,294,086 TERT	T/C	.252- .265	.921-.84	CML	.83 (.78-.88)	1.04 (.95-1.13)	.96 (.87-1.07)	0.37
				UML	.77 (.70-.84)	1.13 (1.02-1.26)	1.10 (.94-1.27)	0.089
				EB	.79 (.73-.87)	1.09 (.98-1.21)	1.04 (.89-1.21)	0.37
5p15.33 rs401681 1,322,087 CLPTM1L	T/C	.466-.463	1-1	CML	1.19 (1.12-1.27)	.98 (.91-1.05)	.97 (.89-1.06)	0.78
				UML	1.20 (1.13-1.28)	.97 (.88-1.06)	.99 (.88-1.12)	0.78
				EB	1.19 (1.12-1.27)	.98 (.91-1.05)	1.00 (.89-1.12)	0.79
7p13 rs17688601 40,866,663 SUGCT	A/C	.252-.259	1-1	CML	.87 (.82-.93)	1.01 (.93-1.09)	1.03 (.93-1.14)	0.82
				UML	.87 (.82-.93)	1.00 (.89-1.11)	1.10 (.96-1.27)	0.31
				EB	.87 (.82-.93)	1.00 (.90-1.10)	1.06 (.94-1.20)	0.55

Abbreviations: Chr:chromosome; MAF:Minor Allele Frequency; GxE: Gene by Environment ; OR 95% CI: Odds Ratio and its 95% confidence interval

[a] SNP position according to NCBI Human Genome Build 37

[b] Quality of imputation metric

Table C.2: Odds Ratios and Interaction P-value GWASs loci in Caucasian population (Contd.)

Chr SNP Position[a] Gene	Effect Allele/ Ref Allele	MAF PanC4 - PanScan	INFO[b] PanC4 - PanScan	Method GxE	OR SNP (95% CI)	OR SNP × FormerSmokers (95% CI)	OR SNP × CurrentSmokers (95% CI)	Interaction P-value
7p12 rs73328514 47,488,569 TNS3	T/A	.109-.111	.965-.929	CML	.83 (.76-.90)	.98 (.88-1.09)	1.01 (.86-1.17)	0.93
				UML	.83 (.74-.92)	1.02 (.87-1.18)	1.04 (.85-1.28)	0.91
				EB	.83 (.76-.90)	1.00 (.88-1.13)	1.01 (.87-1.18)	0.98
7q32.3 rs6971499 130,680,521 LINC-PINT	C/T	.146-.142	1- .954	CML	.84 (.77-.91)	.98 (.89-1.08)	.97 (.86-1.11)	0.89
				UML	.84 (.77-.92)	.99 (.86-1.13)	.97 (.81-1.17)	0.96
				EB	.84 (.77-.91)	.98 (.89-1.09)	.97 (.81-1.17)	0.92
8q21.11 rs2941471 76,470,404 HNF4G	G/A	.419-.418	.997-.997	CML	1.14 (1.07-1.21)	.96 (.90-1.02)	.88 (.81-.96)	0.026
				UML	1.15 (1.08-1.22)	.99 (.90-1.08)	.88 (.77-.99)	0.11
				EB	1.14 (1.07-1.21)	.96 (.90-1.02)	.88 (.81-.96)	0.028
8q24.21 rs10094872 128,719,884 MYC	T/A	.371-.375	.964-.943	CML	1.14 (1.07-1.21)	.95 (.88-1.02)	1.02 (.93-1.12)	0.2
				UML	1.12 (1.06-1.20)	.96 (.87-1.06)	1.09 (.96-1.24)	0.16
				EB	1.14 (1.07-1.21)	.95 (.87-1.03)	1.03 (.93-1.15)	0.23
8q24.21 rs1561927 129,568,078 MIR1208	C/T	.25-.26	1-1	CML	1.12 (1.06-1.19)	.99 (.91-1.08)	.98 (.88-1.09)	0.92
				UML	1.12 (1.05-1.19)	1.00 (.90-1.11)	1.04 (.91-1.20)	0.83
				EB	1.13 (1.06-1.20)	.99 (.90-1.08)	1.01 (.89-1.14)	0.94
9q34 rs505922 136,149,229 ABO	C/T	.373-.366	1-1	CML	1.26 (1.18-1.34)	.98 (.91-1.05)	1.05 (.96-1.16)	0.31
				UML	1.23 (1.16-1.31)	1.02 (.93-1.13)	1.11 (.98-1.27)	0.24
				EB	1.25 (1.18-1.34)	.99 (.91-1.08)	1.07 (.95-1.20)	0.4
13q12.2 rs9581943 28,493,997 PDX1-AS1-PDX1	A/G	.41-.414	1- .987	CML	1.16 (1.09-1.23)	1.01 (.94-1.08)	.98 (.90-1.07)	0.85
				UML	1.15 (1.08-1.22)	1.05 (.96-1.16)	.94 (.83-1.06)	0.19
				EB	1.16 (1.09-1.23)	1.02 (.94-1.11)	.98 (.88-1.08)	0.68

Abbreviations: Chr:chromosome; MAF:Minor Allele Frequency; GxE: Gene by Environment ; OR 95% CI: Odds Ratio and its 95% confidence interval

[a] SNP position according to NCBI Human Genome Build 37

[b] Quality of imputation metric

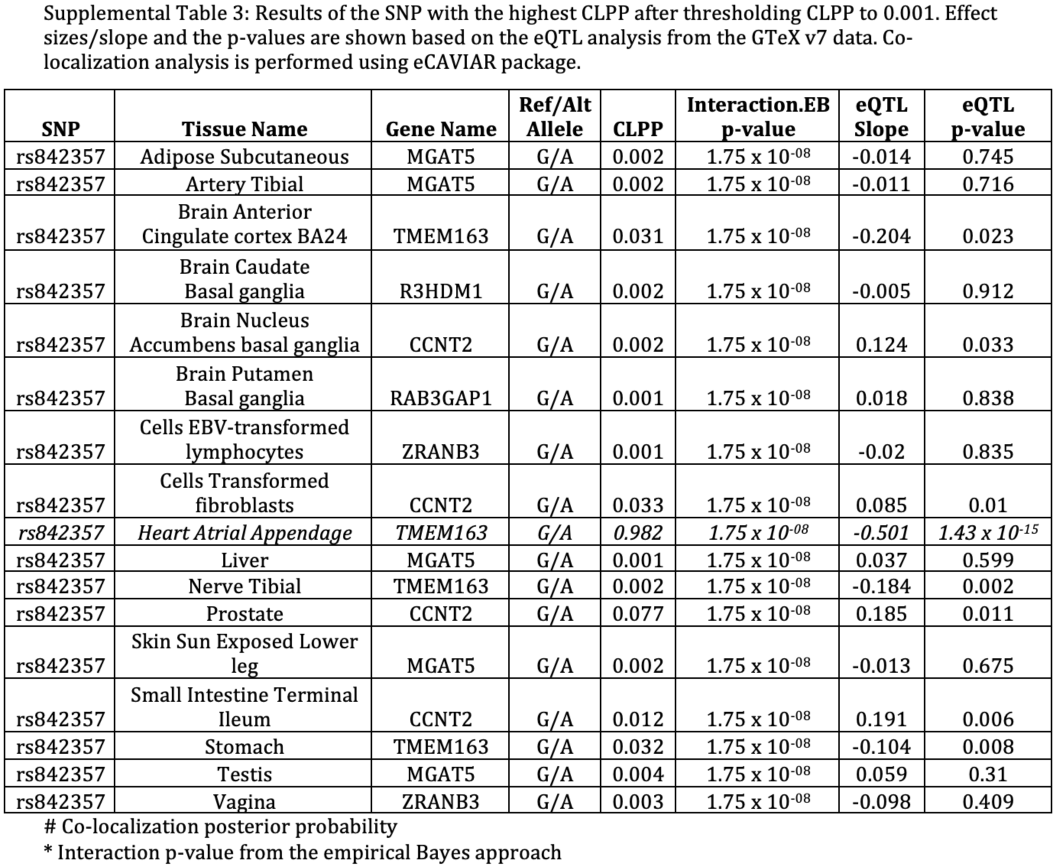
Table C.2: Odds Ratios and Interaction P-value GWASs loci in Caucasian population (Contd.)

Chr SNP Position^a Gene	Effect Allele/ Ref Allele	MAF PanC4 - PanScan	INFO^b PanC4 - PanScan	Method GxE	OR SNP (95% CI)	OR SNP × FormerSmokers (95% CI)	OR SNP × CurrentSmokers (95% CI)	Interaction P-value
13q22.1 rs9543325 73,916,628 KLF5 and KLF12	C/T	.409-.391	1-1	CML	1.29 (1.21-1.37)	.98 (.92-1.05)	.94 (.86-1.03)	0.43
				UML	1.27 (1.20-1.36)	1.04 (.94-1.15)	.91 (.80-1.03)	0.13
				EB	1.28 (1.20-1.36)	1.00 (.92-1.09)	.93 (.84-1.03)	0.33
16q23.1 rs7190458 75,263,661 BCAR1	A/G	.056-.051	1-.739	CML	1.51 (1.32-1.74)	.87 (.74-1.03)	.83 (.67-1.02)	0.13
				UML	1.50 (1.29-1.75)	.86 (.69-1.08)	.88 (.65-1.18)	0.4
				EB	1.51 (1.31-1.73)	.87 (.73-1.04)	.86 (.65-1.14)	0.28
17q12 rs4795218 36,078,510 HNF1B	A/G	.218-.222	.954-.958	CML	.88 (.82-.93)	.97 (.89-1.06)	.99 (.88-1.10)	0.79
				UML	.90 (.83-.99)	.93 (.83-1.05)	.93 (.80-1.08)	0.43
				EB	.88 (.83-.94)	.96 (.88-1.06)	.98 (.85-1.12)	0.73
17q25.1 rs11655237 70,400,166 LINC00673	T/C	.13-.120	.955-1	CML	1.34 (1.23-1.47)	.88 (.79-.98)	.83 (.73-.96)	0.013
				UML	1.32 (1.19-1.47)	.95 (.82-1.09)	.84 (.70-1.02)	0.21
				EB	1.34 (1.23-1.46)	.90 (.79-1.02)	.84 (.73-.97)	0.038
18q21.32 rs1517037 56,878,274 GRP	T/C	.177- .182	1-1	CML	.87 (.82-.93)	1.03 (.94-1.13)	.96 (.86-1.08)	0.48
				UML	.85 (.78-.93)	1.06 (.93-1.20)	1.01 (.87-1.19)	0.66
				EB	.86 (.79-.94)	1.07 (.96-1.19)	.97 (.85-1.11)	0.31
22q12.1 rs16986825 29,300,306 ZNRF3	T/C	.165-.158	1- .996	CML	1.14 (1.04-1.24)	1.00 (.91-1.10)	1.20 (1.07-1.36)	0.0053
				UML	1.18 (1.08-1.28)	.96 (.85-1.10)	1.06 (.90-1.25)	0.05
				EB	1.15 (1.05-1.26)	.99 (.88-1.10)	1.12 (.95-1.32)	0.28

Abbreviations: Chr:chromosome; MAF:Minor Allele Frequency; GxE: Gene by Environment ; OR 95% CI: Odds Ratio and its 95% confidence interval

[a] SNP position according to NCBI Human Genome Build 37

[b] Quality of imputation metric



Supplemental Table 3: Results of the SNP with the highest CLPP after thresholding CLPP to 0.001. Effect sizes/slope and the p-values are shown based on the eQTL analysis from the GTeX v7 data. Co-localization analysis is performed using eCAVIAR package.

SNP	**Tissue Name**	**Gene Name**	**Ref/Alt Allele**	**CLPP**	**Interaction.EB p-value**	**eQTL Slope**	**eQTL p-value**
rs842357	Adipose Subcutaneous	MGAT5	G/A	0.002	1.75 x 10^{-08}	-0.014	0.745
rs842357	Artery Tibial	MGAT5	G/A	0.002	1.75 x 10^{-08}	-0.011	0.716
rs842357	Brain Anterior Cingulate cortex BA24	TMEM163	G/A	0.031	1.75 x 10^{-08}	-0.204	0.023
rs842357	Brain Caudate Basal ganglia	R3HDM1	G/A	0.002	1.75 x 10^{-08}	-0.005	0.912
rs842357	Brain Nucleus Accumbens basal ganglia	CCNT2	G/A	0.002	1.75 x 10^{-08}	0.124	0.033
rs842357	Brain Putamen Basal ganglia	RAB3GAP1	G/A	0.001	1.75 x 10^{-08}	0.018	0.838
rs842357	Cells EBV-transformed lymphocytes	ZRANB3	G/A	0.001	1.75 x 10^{-08}	-0.02	0.835
rs842357	Cells Transformed fibroblasts	CCNT2	G/A	0.033	1.75 x 10^{-08}	0.085	0.01
rs842357	*Heart Atrial Appendage*	*TMEM163*	*G/A*	*0.982*	*1.75 x 10^{-08}*	*-0.501*	*1.43 x 10^{-15}*
rs842357	Liver	MGAT5	G/A	0.001	1.75 x 10^{-08}	0.037	0.599
rs842357	Nerve Tibial	TMEM163	G/A	0.002	1.75 x 10^{-08}	-0.184	0.002
rs842357	Prostate	CCNT2	G/A	0.077	1.75 x 10^{-08}	0.185	0.011
rs842357	Skin Sun Exposed Lower leg	MGAT5	G/A	0.002	1.75 x 10^{-08}	-0.013	0.675
rs842357	Small Intestine Terminal Ileum	CCNT2	G/A	0.012	1.75 x 10^{-08}	0.191	0.006
rs842357	Stomach	TMEM163	G/A	0.032	1.75 x 10^{-08}	-0.104	0.008
rs842357	Testis	MGAT5	G/A	0.004	1.75 x 10^{-08}	0.059	0.31
rs842357	Vagina	ZRANB3	G/A	0.003	1.75 x 10^{-08}	-0.098	0.409

Co-localization posterior probability

* Interaction p-value from the empirical Bayes approach

Figure C.2

Prosenjit Kundu

Curriculum Vitae

Education

2015–Present Pursuing Ph.D. in Biostatistics, The Johns Hopkins University Bloomberg School of Public Health, Advisor: Dr. Nilanjan Chatterjee

2013–2015 M.Stat, Indian Statistical Institute, Kolkata, India

2010–2013 B.Stat(Hons.), Indian Statistical Institute, Kolkata, India

Research Interests

My research focuses on developing statistical methods for integrating multiple data sources with disparate information under different setups induced by the study designs, including two-phase sampling. In particular, I worked on building rich prediction models for breast cancer risk and other types of cancer by combining different models that may have information on a partial set of risk factors. Besides, I work in exploring gene-environmental interactions associated with pancreatic cancer, discovering novel SNPs interacting with exogenous exposures. Most of my methodological interests are, but not limited to, meta-analysis, semi-parametric inference, with the objective of applications in cancer epidemiology and genetics to solve public health problems.

Publications and Software

Publications **Prosenjit Kundu**, Runlong Tang, Nilanjan Chatterjee. Generalized Meta-Analysis for Multiple Regression Models Across Studies with Disparate Covariate Information. *Biometrika*, 2019, 106, 3, 567–585.

Allison Meisner, **Prosenjit Kundu**, Nilanjan Chatterjee, Case-Only Analysis of Gene-Environment Interactions Using Polygenic Risk Scores, *American Journal of Epidemiology*, 2019.

Prosenjit Kundu, and Nilanjan Chatterjee, Analysis of Two-Phase Studies using Generalized Method of Moments, 2019, *arXiv:1910.1199v2 [stat.ME]*.

Software **Prosenjit Kundu**, Runlong Tang, Nilanjan Chatterjee, GENMETA: Implements Generalized Meta-Analysis Using Iterated Reweighted Least Squares Algorithm, R package version 0.1, 2018, *https://cran.r-project.org/package=GENMETA*

Manuscripts in preparation

Prosenjit Kundu*, Evelina Mocci*, et.al., Genome-wide Interaction Scan Identifies Gene by Smoking interaction at 2q21.3 for Pancreatic Cancer Risk. **equal contribution*

The Johns Hopkins University Bloomberg School of Public Health, E3036, Biostatistics
615 N. Wolfe Street., Baltimore, MD 21205
✉ *pkundu@jhu.edu*

Andrew Leroux, Shiyao Xu, **Prosenjit Kundu**, John Muschelli, Ciprian Crainiceanu, Nilanjan Chatterjee, Quantifying the Predictive Performance of Objectively Measured Physical Activity on Mortality in the UK Biobank.

Allison Meisner, **Prosenjit Kundu**, Yan Zhang, Lauren Lan, Sungwon Kim, Disha Ghandwani, Montserrat Garcia-Closas, Nilanjan Chatterjee, Combined Utility of Polygenic Risk Scores across 25 Complex Traits for Predicting Overall Mortality in the UK Biobank.

Awards and Achievements

2019 Young Investigators Award by the ASA Section on Statistics in Epidemiology, Joint Statistical Meetings

2018 Joseph Zeger Travel Reimbursement Award, Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health

2010–2015 Recipient of Innovation in Science Pursuit for Inspired Research (INSPIRE) scholarship by the Department of Science and Technology (DST), Govt. of India.

2008 All India Topper in 10th KVS Junior Mathematics Olympiad.

Presentations

2019 Generalized Meta-Analysis for Combining Disparate Risk Factor Information Across Studies: Inference on Multiple Regression-Based Risk Prediction Models, *Invited Talk, JSM conference.*

2019 Evidence of gene by smoking interaction on 2q21.3 for pancreatic cancer risk, *Invited Talk, Pancreatic Cancer Case-Control Consortium (PanC4), Annual Meeting*

2018 Generalized Meta-Analysis in the Era of Biobanks, *Invited Poster Presentation, The Program in Quantitative Genetics (PQG) Conference on Biobanks: Study Designs and Data Analysis, Harvard School of Public Health*.

2018 Generalized Meta-Analysis: A step towards building rich models by combining information from multiple studies, *Invited Poster Presentation, The Royal Statistical Society (RSS) Conference*.

2017 Overview of UK Biobank: A gold mine for population-based public health research. *Statistical Genetics Working Group, Dept. of Biostatistics, JHSPH*.

Conferences, Meetings and Workshops Attended

2019 The Joint Statistical Meetings (JSM).

2019 Pancreatic Cancer Case-Control Consortium (PanC4), Annual Meeting

2018 The Program in Quantitative Genetics (PQG) Conference, Biobanks: Study Designs and Data Analysis

2018 Scale with HAIL : Genomic Analysis in the Biobank Era, Organized by the HAIL team at the Broad Institute.

2018 The Royal Statistical Society (RSS) International Conference

Responsibilities

2016–Present **UK Biobank data management**, *Key person behind downloading, curating and creating an analytic pipeline for the UK Biobank study at the school/department*, PI of the project is Dr. Nilanjan Chatterjee. .

2016–Present **Teaching Assistant (TA)**, The Johns Hopkins University Bloomberg School of Public Health.

- 140.623.01: Statistical Methods in Public Health III-IV, Fall 2019, Lead TA.
- 140.623.01: Statistical Methods in Public Health III-IV, Spring 2019.
- 140.623.01: Statistical Methods in Public Health I-II, Fall 2018.
- 140.623.01: Statistical Methods in Public Health III-IV, Spring 2018.
- 140.621-622.01: Statistical Methods in Public Health I-II, Fall 2017.
- 140.723-724.01: Probability Theory III-IV, Spring 2017.
- 140.721-722.01: Probability Theory I-II, Fall 2016.

Academic Internships

2014–2015 Analysis of Heart Rate Data under Prof. Ciprian M. Crainiceanu, JHSPH: Longitudinal data on 812 subjects was collected. The data is on heart rates and activity counts. It is a minute by minute data. High and low active bouts have been defined based on some threshold level of the activity counts. For each subject the trend of the heart rate is seen from a high activity period to a low active one. A generalized linear model was fitted to obtain the time coefficient, a measure to study its dependence on age, gender and other variables.

2013–2015 Developed walking characterization tools based on Fast Fourier Transformations(FFT) of the time series for automatic detection of walking from an unlabeled time series of tri-axial acceleration curves under Prof. Ciprian M. Crainiceanu, JHSPH

Technical Skills

Programing Languages	C, C++, HTML, PHP
Software and Tools	R, PLINK, HAIL, QCTOOLS, MATLAB, LaTeX