Statistical Methods on Emerging Medical Studies

by

Youna Hu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Gonçalo R. Abecasis, Co-Chair
Professor Peter X.K. Song, Co-chair
Associate Professor Sebastian Zöllner
Assistant Professor Hyun Min Kang
Assistant Professor Cristen Willer

To Paul, my parents and Addison

ACKNOWLEDGEMENTS

This dissertation is made possible with so many great people's guidance and help.

I would like to thank Dr. Peter Song, who has introduced me to the world of statistics. I came to this department with a mathematics background and he met with me frequently to work on statistical application in medical research so that I could smoothly transit from mathematics to statistics. His rigid statistical thinking, his boundless knowledge and his infinite patience have helped me to enter the world of statistics with great joy.

I also would like to say thank you to Dr. Gonçalo Abecasis, who accepted me into his lab which pioneers the world-wide statistical genetics. He gave me the chance to work on the project with GlaxoSmithKline on incorporating genetics into clinical trials. He also sent me to Wellcome Trust summer school, where I opened my eyes by so many interesting talks and interactions with world leading scientists. Moreover, he has encouraged me to attend all the Exome Sequencing Project in person meeting in DC so that I have collected the first-hand experience of the cutting edge exome sequencing studies. I feel so indebted to his encouragement and his generosity in providing me with all these opportunities. In addition, I am mostly grateful to him for teaching me how to think, write and contribute to the scientific community. I feel very fortunate to be able to witness and learn some

of his amazing skills – from his quick intuition, powerful computer code to his efficient scientific writing style.

I also owe my gratitude to Dr. Hyun Min Kang, whose computational shrewdness and dedication have been a unbelievable tale among students. He introduced me to Unix system, scripting, and all the useful computation skills. He also provided so many valuable tools so I could work with the vast amount of genetic data. In every meeting that I had with him, no matter how short it is, I would always learn some new computational tricks and skills. I've also learnt so much statistics genetics from him by working with him on the clinical trial enrichment and ancestry inference projects.

I would also like to express my gratefulness to Dr. Cristen Willer and Dr. Sebastian Zöllner, who have inspired me in various ways. Working with Dr. Willer in the exome sequencing project is a great venue for me to learn genetics and applied statistics as fast as possible. Her kindness in including me to her lab meeting, in teaching me biology and in sharing her wisdoms in various things will always be memorable. She is also a role model in beautifully balancing career and family. Dr. Zöllner is another great mentor to me, either by instantaneously writing down the most detailed equations to explain population genetics to me or sharing his opinions on my academic aspiration.

Next, I would like to thank all my family members. My husband, Paul Ullrich is not only a great companion in daily life but also can be very helpful in the academic world. He has showed me the patience and the meticulousness required in computer programming. He has also been a great person to discuss mathematical equations with. It was perfect to have him around. My parents have never stopped encouraging and supporting me in pursuing my academic goal. They have alway

pushed me to work hard, to embrace challenges and to achieve my best. My son Addison has also served as an important role for this thesis. He has "trained" me to be calmer and more patient, which is also very important in writing this dissertation.

Then, I would like to thank everyone in the Abecasis group, Center for Statistical Genetics and Department of Biostatistics, especially Dr. Mike Boehnke, Xiaowei Zhan, Dajiang Liu, Goo Jun, Yancy Lo, Carlo Sidore, Shuang Feng, Xuejing Wang, Meihua Wu and Laura Baker. All the interaction and discussion that I had with people here will be the most memorable in my life time.

In the end, I would like to thank Chaolong Wang for carefully reading parts of my dissertation and give valuable suggestions.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Statistical Methods on Emerging Medical Studies

by

Youna Hu

Co-Chair: Gonçalo R. Abecasis, Peter X.K. Song

Many areas of medical research can benefit from the application of new statistical methods. In this dissertation, we demonstrate the possibilities in some emerging medical studies.

In chapter 2, we describe the design of longitudinal trials with dichotomized outcomes. Longitudinal studies are often analyzed using Generalized Estimating Equations (GEE) or Quadratic Inference Functions (QIF). We explore QIF-based analysis for study design: (i) we derive and compared sample size and power for QIF and GEE; (ii) we propose an optimal scheme of sample size determination to overcome the difficulty of unknown true correlation matrix; and (iii) we show that QIF based analysis become more efficient as the number of follow-up visits increases. We illustrate that without sacrificing power, the QIF design leads to sample size an study cost savings than the GEE analysis.

In chapter 3, we focus on the analysis of exome sequencing studies. These studies focus sequencing resources on the exome but often yield many reads outside

the targeted regions. These reads are almost always discarded but we propose our method SEQMIX to incorporate genotype likelihoods of off-targeted reads and show that it can decipher the local ancestry with resolution similar to that obtained with modern genome-wide association studies (GWAS) arrays. Our estimates are useful for analysis of human population history and disease gene mapping studies with exome and targeted sequence data.

In chapter 4, we develop a general framework to explore benefits of incorporating genetic risk into prevention trial designs. We consider screening, recruitment and follow-up costs and current genetic findings in diseases. We consider type 1 diabetes (T1D), type 2 diabetes (T2D), myocardial infarction (MI) and age-related macular degeneration (AMD) as examples as they have distinct genetic architecture: many small effect markers are associated with T2D and MI; whilst loci with large effect size have been identified for T1D and AMD. We quantify the benefits by illustrating settings where reduction of trial cost or duration is up to 70% and settings where savings are modest. The benefits depend on the disease genetic architecture, but we also project that benefits will increase.

CHAPTER I

Introduction

Many areas of medical research can benefit from the advancement of new statistical methods. In this dissertation, we focus statistical applications on clinical trials design and analysis of whole exome sequencing studies. Here, we give an overview of efficient clinical trials, whole exome sequencing studies and local ancestry inference for admixed individuals. We also briefly explain the connections between genetic studies and clinical trials.

1.1 Efficient clinical trials

Clinical trials are an important approach undertaken in medical research and drug development. A typical trial often recruits a large number of participants, spans a long period of time and requires extensive resources. Typically, the participants are randomized into two or more treatment arms and followed for 5 – 10 years. Various biological measurements are obtained at baseline as well as throughout the trial. In the end, the efficacy and safety of the treatments are determined by statistical analysis of the collected measurements from the trials.

Clinical trials date back as far as 1750 and since then have become the norm of modern evidence-based medicine. The National Institute of Health (NIH) (`www.clinicaltrials.gov`) categorizes trials into these six different types: Prevention trials, screening trials, diagnostic trials, treatment trials, quality of life trials and compassionate use trials. In addition, clinical trials are composed of various phases: Pre-clinical studies, phase 0 and phase I to phase IV. Clinical trials also include many stages: design, recruitment, randomization, interim analysis, termination of trial, final data analysis and result summary and report.

Statistical methods play important roles in various stages of the clinical trial, from design, data analysis, to the interpretation of the results. At the design stage, to reduce the variability of its implementation, a protocol is prepared to describe the sample size, statistical analysis method, hypothesis as well as other procedures. An efficient clinical trial should be planned as early as this stage. Such a trial requires a smaller sample size without losing the statistical power, demands a shorter amount of time for following up participants and ultimately, can lead to savings in cost and resources. An efficient clinical trial, from an ethical point of view, exposes fewer participants to the risk of new drugs or the relative inefficiency of the comparative treatment, speeds up the mass production and application of effective drugs and hence can reduce suffering of the patients. Efficient clinical trial can be achieved by incorporating new statistical methods into the trial.

In this dissertation, we develop statistical methods to improve the efficiency of a treatment trial and prevention trial.

1.2 Whole exome sequencing studies

Whole exome sequencing is the "sweet spot" before whole genome sequencing (*Teer and Mullikin*, 2010) as it only targets the 1% of the whole genome which is responsible for protein coding and harbors a majority (85% by *Choi et al.* 2009) of the mutations with large effects on various diseases.

Exome sequencing is an effective tool for Mendelian disease gene discovery (*Bamshad et al.* 2011). The first few successful examples include Freeman-Sheldon Syndrome (*Ng et al.* 2009), Miller syndrome (*Ng et al.* 2010b) and Kabuki syndrome (*Ng et al.* 2010a). Since then, an emerging number of exome sequencing studies are being undertaken for various rare Mendelian traits (ASHG 2011 abstracts).

Exome sequencing has also been shown to assist in accurate medical diagnosis. *Choi et al.* (2009) illustrated a clinical story in changing the initial misdiagnosis of Bartter syndrome, a renal salt-wasting disease, into the diagnosis of congenital cholride diarrhea. This correction was based on the exome sequencing finding of a homozygous missense $D652N$ mutation at a position in $SLC26A3$ and was later confirmed in further clinical diagnosis. Moreover, *Worthey et al.* (2011) provided another example of using exome sequencing to diagnose a child with intractable inflammatory bowel disease. Other examples include diagnosis of cases in Charcot-Marie-Tooth neuropathy (*Lupski et al.*, 2011), severer brain malformations (*Bilguvar et al.*, 2011) and so on.

Exome sequencing has a potential for complex disease gene mapping to find rare variants (minor allele frequency less than 5%) that were previously missed by genome-wide association studies (GWAS). Although many common variants

have been identified for many highly heritable diseases, these significant loci only explain a small portion of the heritability (*Maher* 2008). A well accepted idea is that rare variants contribute to a big portion of the missing heritability (*Manolio et al.* 2009). The study of rare variants from exome sequencing shows promise in understanding the missing heritability (*Cirulli and Goldstein* 2010).

Whole exome sequencing focuses on the protein-coding regions and the remaining 99% of the genome falls out of scope. Consequently, it looks like that exome sequencing experiments cannot answer scientific questions that requires data beyond coding regions. However, *Pasaniuc et al.* (2012) showed that extremely low coverage off target reads can be used for genome-wide genotype imputation. In this dissertation, we propose a method to use off target reads to infer local ancestry across the entire genome.

1.3 Local ancestry inference for admixed individuals

An admixed individual is one whose genome is a mosaic of genomes from separated populations. Initially, two parents from population 1 and 2 come together and produce the first-generation admixed offspring. This generation inherits one chromosome from each parent and their ancestry is uniform across either chromosome. Starting from the second generation, the uniformity is broken down and the admixed genome is composed by alternating parental genomes along the genome. The ancestry for the genomic pieces are what we refer to as local ancestry.

Deciphering the ancestry of genomic pieces in admixed individuals is important for studies of human evolutionary history (*Xu et al.* 2008; *Gravel* 2012). By understanding the size of the ancestry blocks, we can infer how many years ago

that the ancestral populations were mixed and in what ratio. Local ancestry estimates are also important in disease gene mapping studies (*Shriner et al.* 2011). They could be used for controlling population stratification in associating disease with genotypes from admixed populations. Because admixed individuals do not necessary have the same proportion of ancestry from each parental population, the allele frequency, linkage disequilibrium and population diversity at each locus are different than the parent populations. Population stratification is an important question to address in disease mapping studies of such populations.

Previously, this ancestry deconvolution relied on the availability of genome-wide genotypes, such as from high-density GWAS arrays. On such data, one could use methods that requires a step to prepare for ancestry informative markers (AIMs) (*Patterson* 2004); one could divide whole genome into many windows and then model each window as a unit (*Sankararaman et al.* 2008; *Bryc et al.* 2010a); one could use methods that model local linkage equilibrium (*Tang et al.* 2006); or even, one could apply haplotype based methods such as HAPMIX (*Price et al.* 2009), HAPAA (*Sundquist et al.* 2008). These methods have been applied for admixed populations such as Afrian American (*Bryc et al.* 2010a), Latinos (*Bryc et al.* 2010b) and Uyghur (*Xu et al.* 2008) in China.

In this thesis, we focus on local ancestry inference for exome sequenced samples for whom we do not have genome-wide high density genotype data. We show that inferred ancestry using our methods is very similar to that derived using the genotype array data.

1.4 Genetics in clinical trials

At first appearance, genetic studies and clinical trials answer distinct scientific questions, but in fact they are connected in various ways. Genetic studies have led to many novel discoveries in modern medical research. Some of these discoveries are so important that they have quickly become adopted into clinical trials. For example, there has been exploration of BRCA1 and BRCA2 variants in breast cancer studies (*King et al.* 2001) and untilisation of ApoE/e4 variants for Alzheimer's disease (*Cummings et al.* 2007). Recently, researchers have proposed to incorporate genetic information into clinical trials (*Simon* 2008).

Genetic information is critical in pharmacogenomics/pharmacogenetics studies which are conducted to understand the impact of genetic variations to differences in drug (or treatment) response and adverse effect. Since genetic component is an important part in metabolic pathways, patients of different genetic variants react differently to the type or the dose of drugs. Understanding the relation between metabolic function and genetics makes it possible to understand the drug response in different populations of distinct genetic makeup. One example of such finding is the mechanism of how the enzyme thiopurine methyltransferance (TPMT) responds to drugs in childhood leukemia and autoimmune diseases. Another breakthrough is the discovery of genetic polymorphisms near the human IL28B gene that affects the effectiveness of a type of Hepatitis C treatment.

In this dissertation, we describe a statistical method that provides a unified framework in using genetic information in clinical trial designs. We provide a systematic and quantitative approach to incorporate genetic findings in the prevention trials of diseases.

1.5 Outline of dissertation

In this dissertation, we describe statistical methods for efficient clinical trials and accurate inference of local ancestry for exome sequenced individuals. This dissertation is organized as follows. In Chapter 1, we give an overview of efficient clinical trials, next-generation whole exome sequencing studies and local ancestry inference for admixed individuals. Chapter 2 describes the sample size determination for quandratic inference functions (QIF) in longitudinal design with dichotomous outcomes and recommends QIF to be considered at the clinical trial design stage. Chapter 3 focuses on using off target reads from exome sequence data to accurately infer the local ancestry for admixed individuals at the whole genome level. Chapter 4 combines these two distinct areas of research to illustrate how genetic findings may improve current clinical trial designs.

CHAPTER II

Sample Size Determination for Quadratic Inference Functions in Longitudinal Design with Dichotomous Outcomes

2.1 Introduction

Longitudinal clinical studies are undertaken extensively in biomedical sciences. In a longitudinal clinical study, repeated measurements are recorded at pre-scheduled time points during a study period. One primary objective of a clinical study is to compare effects of test treatments with those of controlled treatments.

When outcome variables of interest are binary, the marginal logistic model is popular for assessing the population-average effect of a test treatment. For the details regarding marginal generalized linear models for longitudinal data, refer to *Diggle et al.* (2002) and *Song* (2007). The method of generalized estimating equations (GEE), proposed by *Liang and Zeger* (1986), has been widely applied to estimate and infer treatment effects in the marginal logistic model. Accordingly, sample size determination in the GEE-based design has been studied in

the literature; for example *Diggle et al.* (2002) (chapter 2), *Pan* (2001a), *Jung and Ahn* (2003) and *Rochon* (1998), among others. It is worth pointing out that *Diggle et al.* (2002), *Pan* (2001a) and *Rochon* (1998) considered a very simple longitudinal model that contains only a single treatment covariate; neither time covariate nor treatment-time interaction covariate is included in the design. *Jung and Ahn* (2003) considered all these covariates but they treated the working correlation structure as independent. Although these existing tools have addressed basic needs in the longitudinal study design, sample size and power calculation under more general scenarios, say, longitudinal models with both covariates of time and treatment-time interaction with various working correlations considered, are clearly of great importance in clinical research.

One objective of this paper is to establish a new scheme of sample size and power calculation based on the quadratic inference functions (QIF) approach (*Qu et al.*, 2000). QIF is a powerful alternative method of estimation and inference to the popular GEE. It has been shown in the literature that QIF is superior to the GEE with respect to efficiency gain and robustness against outliers (*Song et al.*, 2009). Our sample size and power calculation will be based on the Wald test in marginal logistic models. In the literature, the Wald test based design in the logistic model for cross-sectional data has been investigated by *Demidenko* (2006) and *Hsieh et al.* (1998), among others. Unfortunately, these existing formulas cannot be easily modified to suit longitudinal studies due to the presence of within subject correlation. In addition, we compare QIF and GEE in terms of their sample size and power. We also examine how the sample size is affected by varying the cluster size (i.e. the number of repeated measurements or follow-up visits per subject). As shown in our numerical examples, the GEE sample size may escalate

substantially with an increased number of follow-up visits, whereas the QIF sample size remains nearly unchanged.

In the longitudinal design, a fundamental obstacle is the lack of knowledge about the true correlation matrix. Although in practice one correlation structure (e.g. compound symmetry) can be used to determine the sample size, the result is then subject to risk of either over- or under-estimation of the needed sample size. In particular, once the data are collected, the true correlation structure can be obtained, which is very likely to be different from the one used initially in the design. To overcome this difficulty, we propose an optimal strategy of sample size determination from a perspective of minimal average risk. We postulate that the true correlation matrix is an element of a class of correlation matrices that are governed by a certain prior Wishart distribution. The hyper-parameter in the prior distribution may be a correlation matrix specified according to previously acquired knowledge from a pilot study, from similar studies in the literature or simply from an assumption of non-informative independent correlation. Then, we obtain the optimal sample size that is derived by minimizing the average risk under such prior distribution.

This article is organized as follows. Section 2 presents a brief introduction to the QIF method. Section 3 is devoted to the sample size determination for the Wald test in the marginal logistic model in both QIF- and GEE-based designs, in which we present an optimal strategy of sample size determination. Section 4 investigates, analytically and numerically, the relationship between sample size and cluster size (i.e., the number of follow-up visits). Section 5 contains some concluding remarks. Technical details and information regarding the R package QIFSAMS are included in the appendices.

2.2 Quadratic Inference Functions

In a balanced longitudinal clinical trial, y_{ij} denotes the outcome of subject i at time point t_j. There are n subjects in the study, and m repeated measurements planned to be collected from each of the n subjects. Thus the total number of observations is $N = n \times m$. We further assume that the observations from different subjects are independent and those of the same subject are correlated. Both GEE and QIF methods postulate that the marginal mean, μ_{ij}, of the outcome y_{ij}, is a function of some covariates through a link function g, namely $g(\boldsymbol{\mu}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the regression coefficient. Here $\mathbf{x}'$ denotes the transpose of matrix $\mathbf{x}$. The variance of y_{ij} is a function of the mean $\text{var}(y_{ij}) = \phi V(\mu_{ij})$, where ϕ is the dispersion parameter. Vectors $\mathbf{y}_i$ and $\boldsymbol{\mu}_i$, with elements y_{ij} and μ_{ij} respectively, denote the longitudinal measurements and their mean for subject i. To obtain an estimate of $\boldsymbol{\beta}$, the GEE method solves

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i{}' \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.1}$$

where $\dot{\boldsymbol{\mu}}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is an $n \times p$ matrix, and $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\rho})\mathbf{A}_i^{1/2}$ with $\mathbf{A}_i$ being the diagonal matrix of the marginal variances, $\text{var}(y_{ij})$, and $\mathbf{R}_i(\boldsymbol{\rho})$ being the working correlation matrix.

The QIF method, proposed by *Qu et al.* (2000), is derived from the fact that the inverse of the working correlation matrix $\mathbf{R}(\boldsymbol{\rho})$ can be approximated by a linear combination of several basis matrices:

$$\mathbf{R}(\boldsymbol{\rho})^{-1} \approx \sum_{l=0}^{k} a_l(\boldsymbol{\rho})\mathbf{M}_l, \tag{2.2}$$

where $\mathbf{M}_0 = \mathbf{I}$ is the identity matrix, $\mathbf{M}_1, \ldots, \mathbf{M}_k$ are known basis matrices with entries 0 or 1 and $a_0(\boldsymbol{\rho}), \ldots, a_k(\boldsymbol{\rho})$ are unknown coefficients depending on the parameter $\boldsymbol{\rho}$. Expression (2.2) holds exactly for some commonly used working correlation structures. For example, the compound symmetry (CS) correlation structure corresponds to $\mathbf{R}(\boldsymbol{\rho})^{-1} = a_0(\boldsymbol{\rho})\mathbf{I} + a_1(\boldsymbol{\rho})\mathbf{M}_1^{cs}$, where the entries of $\mathbf{M}_1$ are 0 along the diagonal and 1 elsewhere. And the AR-1 correlation structure can be written as $\mathbf{R}(\boldsymbol{\rho})^{-1} = a_0(\boldsymbol{\rho})\mathbf{I} + a_1(\boldsymbol{\rho})\mathbf{M}_1^{ar1} + a_2(\boldsymbol{\rho})\mathbf{M}_2^{ar1}$, where $\mathbf{M}_1$ has 1 on the two main off-diagonals and 0 elsewhere and $\mathbf{M}_2$ has 1 on the two corner components of the diagonal.

Plugging expression (2.2) into (2.1) leads to a linear combination of the elements of an extended score vector

$$\overline{\mathbf{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}) \approx \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n}\dot{\boldsymbol{\mu}}_i{}'\mathbf{A}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \sum_{i=1}^{n}\dot{\boldsymbol{\mu}}_i{}'\mathbf{A}_i^{-1/2}\mathbf{M}_k\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}. \tag{2.3}$$

Then, the GEE (2.1) is expressed as $n\mathbf{a}'(\boldsymbol{\rho})\overline{\mathbf{g}}_n(\boldsymbol{\beta})$ where $\mathbf{a}(\boldsymbol{\rho}) = (a_0(\boldsymbol{\rho}), a_1(\boldsymbol{\rho}), \ldots, a_k(\boldsymbol{\rho}))'$ is the vector of the coefficients in expansion (2.2). Since there are more equations than unknown parameters in (2.3), the generalized method of moments proposed by *Hansen* (1982) is then applied to minimize the following quadratic inference function:

$$Q_n(\boldsymbol{\beta}) = n\overline{\mathbf{g}}_n'(\boldsymbol{\beta})\mathbf{C}_n^{-1}(\boldsymbol{\beta})\overline{\mathbf{g}}_n(\boldsymbol{\beta}), \tag{2.4}$$

where $\mathbf{C}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i'(\boldsymbol{\beta})$ is the sample covariance matrix of $\bar{\mathbf{g}}_n$. It is noted that the objective function given in (2.4) only contains $\boldsymbol{\beta}$, and only the basis matrices from the working correlation structure are used to formulate this

function. This implies that the QIF estimator, $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta})$, is obtained without estimating the nuisance parameter $\boldsymbol{\rho}$. This property delivers substantial ease in the design setting since the QIF sample size determination does not require knowledge of parameter $\boldsymbol{\rho}$.

According to *Qu et al.* (2000), the QIF estimator $\hat{\boldsymbol{\beta}}$ has the usual large sample properties; for example, it is $\sqrt{n}$-consistent and asymptotically normal. The asymptotic variance is given by the inverse of the Godambe information matrix (or the sandwich estimator) with consistent estimate $\mathbf{J}_n(\hat{\boldsymbol{\beta}})^{-1}$, with $\mathbf{J}_n(\hat{\boldsymbol{\beta}}) = \bar{\dot{\mathbf{g}}}_n(\hat{\boldsymbol{\beta}})'\mathbf{C}_n^{-1}(\hat{\boldsymbol{\beta}})\bar{\dot{\mathbf{g}}}_n(\hat{\boldsymbol{\beta}})$. The asymptotic normality allows us to establish the Wald test for the hypothesis with respect to the regression coefficients. Again, we note that this asymptotic covariance matrix $\mathbf{J}_n$ is not dependent on the parameter $\boldsymbol{\rho}$. Similarly, the Wald test also applies to the GEE method, in which the asymptotic covariance matrix actually depends on the correlation parameter $\boldsymbol{\rho}$ explicitly. Therefore, our following derivation and comparison of sample size and power will focus on the Wald test that is available in both QIF and GEE.

This QIF method has been coded both in SAS and R language. Both packages can be download at `http://www-personal.umich.edu/~pxsong/`.

2.3 Sample size determination

In this section, we derive sample size n under a fixed m. We begin by describing our model, and then present steps related to the derivation of sample size in both QIF and GEE. Two examples of designs based on the CS and AR-1 working correlation structures are discussed in detail.

2.3.1 Longitudinal logistic model

We consider the following logistic model with longitudinal dichotomous outcomes:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 d_i + \beta_2 t_j + \beta_3 d_i t_j, \tag{2.5}$$

where $\mu_{ij} = P(y_{ij} = 1|d_i, t_j)$ is the probability of a favorable clinical outcome ($y_{ij} = 1$) at visit time t_j for subject i, covariate d_i is the indicator of treatment group, defined as

$$d_i = \begin{cases} 1, & \text{if subject } i \text{ is in test treatment (Rx) group,} \\ 0, & \text{if subject } i \text{ is in controlled treatment group,} \end{cases}$$

and covariate t_j is the time of the j-th visit for subject i. It follows from model (2.5) that the design matrix is $\mathbf{X}_i = (\mathbf{1}, d_i\mathbf{1}, \mathbf{t}, d_i\mathbf{t})'$, where $\mathbf{1}$ is an m-element vector of all ones, $\mathbf{t} = (t_1, t_2, \cdots, t_m)'$, and the vector of regression coefficients is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$. Also, in the logistic model, the variance function is $V_{ij} = \text{V}(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$, and the dispersion parameter $\phi = 1$. Note that under the balanced design with homogeneous visit times, there are only two versions of design matrices, and so are their induced matrices, corresponding to the two respective treatment arms. In all the subsequent expressions below, a sub-index 1 (or 0) denotes terms from the test (or controlled) treatment arm. For example, the design matrix for the test drug arm is $\mathbf{X}_1$ and the counterpart for the controlled drug arm is $\mathbf{X}_0$.

We are interested in testing for one of these three hypotheses: total effect, $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$; main Rx effect, $H_0 : \beta_1 = 0$; and joint Rx effect,

$H_0 : \beta_1 = \beta_3 = 0$. In general, we may express the three scenarios in a unified form as $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}_0 \neq \mathbf{0}$, where $\mathbf{H}$ is a suitable matrix determined by the null hypothesis.

Let $\mathbf{J}_n(\boldsymbol{\beta})$ be the asymptotic covariance matrix of either the QIF estimator or the GEE estimator and let λ_n be the non-centrality parameter given by $\lambda_n = \mathbf{h}_0'\{\mathbf{H}\mathbf{J}_n(\hat{\boldsymbol{\beta}})^{-1}\mathbf{H}'\}^{-1}\mathbf{h}_0$. Then, the Wald test statistic is $(\mathbf{H}\hat{\boldsymbol{\beta}})'\{\mathbf{H}\mathbf{J}_n(\hat{\boldsymbol{\beta}})^{-1}\mathbf{H}'\}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}})$, and the power is given by $1-\eta = \int_{\chi^2_{rk(\mathbf{H})}(1-\alpha)}^{\infty} f(x, \mathrm{rk}(\mathbf{H}), \lambda)dx$, where α is the type I error, η is the type II error, and $f(x, \mathrm{rk}(\mathbf{H}), \lambda_n)$ is the non-central chi-square density function with degrees of freedom $\mathrm{rk}(\mathbf{H})$. Ultimately, we need to find the smallest n such that $1 - \eta \leq \int_{\chi^2_{rk(\mathbf{H})}(1-\alpha)}^{\infty} f(x, \mathrm{rk}(\mathbf{H}), \lambda_n)dx$. Numerically, it is obtained by the forward search algorithm. Notably, the sample size calculation we provided is based on asymptotic properties of the two methods. Teerenstra et al. *Teerenstra et al.* (2010) have considered a small sample setting for GEE sample size. Similar setting for QIF needs separate investigation.

2.3.2 QIF sample size under unknown true correlation

Let us begin with an ideal scenario where the true correlation structure is known. This rather idealized condition will be removed in the next section. We focus on two important cases of designs, respectively under the CS and AR-1 working correlation structures, which are widely used in practice. For both working correlations, we present the details of calculating the asymptotic covariance matrix $\mathbf{J}_n$ for the QIF method, and leave the calculation of this matrix for the GEE to Appendix B.

Design under CS Working Correlation. As shown in section 2.2, the CS structure leads to two basis matrices. Under model (2.5), we explicitly derive the

elements of the extended score vector in (2.3) to be

$$\mathbf{g}_i = (\mathbf{1}', d_i\mathbf{1}', \mathbf{t}_i', d_i\mathbf{t}_i', \boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}, d_i\boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}, \tilde{\boldsymbol{\omega}}_i'\mathbf{A}_i^{-\frac{1}{2}}, d_i\tilde{\boldsymbol{\omega}}_i'\mathbf{A}_i^{-\frac{1}{2}})'\mathbf{e}_i \stackrel{def}{=} \mathbf{B}_i\mathbf{e}_i, \qquad (2.6)$$

where $\mathbf{e}_i = (y_{i1} - \mu_{i1}, y_{i2} - \mu_{i2}, \cdots, y_{im} - \mu_{im})'$, $\mathbf{A}_i = \text{diag}(V_{i1}, \cdots, V_{im})$ and

$$\begin{aligned}
\boldsymbol{\omega}_i' &= \sum_{j=1}^{m} \mathrm{V}(\mu_{ij})^{\frac{1}{2}}\mathbf{1}' - [\mathrm{V}(\mu_{i1})^{\frac{1}{2}}, \cdots, \mathrm{V}(\mu_{im})^{\frac{1}{2}}], \\
\tilde{\boldsymbol{\omega}}_i' &= \sum_{j=1}^{m} t_j\mathrm{V}(\mu_{ij})^{\frac{1}{2}}\mathbf{1}' - [t_1\mathrm{V}(\mu_{i1})^{\frac{1}{2}}, \cdots, t_m\mathrm{V}(\mu_{im})^{\frac{1}{2}}].
\end{aligned}$$

Note that there are only two versions of $\mathbf{A}_i$'s, $\boldsymbol{\omega}_i$'s, $\tilde{\boldsymbol{\omega}}_i$'s as explained in section 2.3.1. Now denote $\bar{d}$ as the proportion of subjects assigned to the test treatment arm. Then, there are $\bar{d}n = \sum_{i=1}^{n} d_i$ and $(1-\bar{d})n$ subjects in the test and controlled treatment arms, respectively. Denoting the true correlation matrix as $\mathbf{R}_T$, we obtain

$$\begin{aligned}
\mathbf{C}_n(\boldsymbol{\beta}) &= \bar{d}\mathbf{B}_1\mathbf{A}_1^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_1^{\frac{1}{2}}\mathbf{B}_1 + (1-\bar{d})\mathbf{B}_0\mathbf{A}_0^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_0^{\frac{1}{2}}\mathbf{B}_0, \qquad (2.7) \\
\dot{\bar{\mathbf{g}}}_n &= -\left\{\bar{d}\mathbf{B}_1\mathbf{A}_1\mathbf{X}_1' + (1-\bar{d})\mathbf{B}_0\mathbf{A}_0\mathbf{X}_0'\right\}, \qquad (2.8)
\end{aligned}$$

which are then used to calculate Godambe information matrix

$$\mathbf{J}_n(\boldsymbol{\beta}) = n\dot{\bar{\mathbf{g}}}_n'(\boldsymbol{\beta})\mathbf{C}_n^{-1}(\boldsymbol{\beta})\dot{\bar{\mathbf{g}}}_n(\boldsymbol{\beta}).$$

Design under AR-1 Working Correlation. For the AR-1 structure there are three basis matrices, as stated in section 2.2. According to *Qu et al.* (2000), the third matrix $\mathbf{M}_2$ makes little contribution to the QIF, and hence is omitted in the

formulation of QIF for convenience. Consequently, we obtain

$$\tilde{\mathbf{g}}_i = (\mathbf{1}, d_i\mathbf{1}, \mathbf{t}_i, d_i\mathbf{t}_i, \mathbf{X}_i\mathbf{A}_i^{\frac{1}{2}}\mathbf{M}_1\mathbf{A}_i^{-\frac{1}{2}})'\mathbf{e}_i \stackrel{def}{=} \tilde{\mathbf{B}}_i\mathbf{e}_i. \tag{2.9}$$

The Godambe information matrix is obtained in the same manner as in the first case of CS correlation, in which $\mathbf{B}_i$ is replaced by $\tilde{\mathbf{B}}_i$ in (2.7) and (2.8).

In both cases above, we also derived the sensitivity matrix $\mathbf{S}_n(\boldsymbol{\beta})$ and the variability matrix $\mathbf{W}_n(\boldsymbol{\beta})$ in the GEE context, and hence the Godambe information matrix of the GEE estimator is given by $\mathbf{J}_n(\boldsymbol{\beta}) = \mathbf{S}_n(\boldsymbol{\beta})'\mathbf{W}_n(\boldsymbol{\beta})^{-1}\mathbf{S}_n(\boldsymbol{\beta})$. See Appendix B for details regarding the $\mathbf{S}_n$ and $\mathbf{W}_n$ matrices.

It is worth pointing out that for model (2.5) with either CS or AR-1 correlation structure, the weight matrix $\mathbf{C}_n$ in the QIF method is singular. Hence, we delete linearly dependent elements in the extended score vector. Since such elements do not present new information, it is sensible to not use them. In fact, when model (2.5) only includes the treatment effect and the working correlation is CS, QIF degenerates to QIF, exactly as obtained by reducing these dependent elements, the QIF method reduces to GEE. Related details can be found in Appendix A. Also, we again emphasize in QIF the working correlation structure contributes the two basis matrices, not the value of parameter $\boldsymbol{\rho}$, to the calculation of matrix $\mathbf{J}_n$. This is not the case for the GEE, where the matrix $\mathbf{J}_n$ depends on the entire $\mathbf{R}_W$ matrix, including the actual value of $\boldsymbol{\rho}$.

2.3.3 Optimal sample size with unknown true correlation

Let R_T denote the true correlation matrix. So, the actual sample size would be $n(R_T)$ if the R_T were known. To overcome the difficulty of an unknown true

correlation structure in the sample size determination, we propose to vary the underlying correlation matrix among a class of possible candidates according to a Wishart prior distribution, instead of fixing it to be a pre-chosen single correlation matrix. This will allow us to reduce the subjectivity related to the choice of the true correlation matrix in the design. To be precise, let correlation matrix $\mathbf{R} \sim \text{Wishart}(\cdot|\mathbf{R}_0)$, where $\mathbf{R}_0$ is a pre-specified correlation matrix. Then, for each sampled correlation matrix $\mathbf{R}$, the sample size $n = n(\mathbf{R})$ is obtained by the sample size determination procedure in section 3.2. We aim to choose an optimal sample size that minimizes the following average risk:

$$E\{n(\mathbf{R}) - n(\mathbf{R}_T)\}^2 = \int_{\mathbf{R}\in\mathcal{R}} \{n(\mathbf{R}) - n(\mathbf{R}_T)\}^2 g(\mathbf{R}) d\mathbf{R},$$

where $g(\mathbf{R})$ is the density of Wishart distribution and $\mathcal{R}$ is the space of correlation matrices. It is easy to see that the optimal sample size is $n^* = E_{\mathbf{R}} n(\mathbf{R}) = \int_{\mathbf{R}\in\mathcal{R}} n(\mathbf{R}) g(\mathbf{R}) d\mathbf{R}$. Practically, this integral can be evaluated by the Monte Carlo simulation method. In effect, the Monte Carlo simulation will bring in variation of correlation matrices, and as a result, we can obtain not only the mean sample size n^*, but also a sample size distribution which in practice, may be more valuable as it provides more options to practitioners. Moreover, this procedure gives us a venue where we can conduct a fair comparison between QIF and GEE analysis, because in this case the comparison will not depend on specific choices of true correlation structures.

2.3.4 Numerical illustration

In this section, we provide three examples to illustrate the proposed QIF and GEE sample size formulas and the comparison of these two methods, based on known and unknown true correlation structures, respectively.

Example 1: Consider several different combinations of true ($\mathbf{R}_T$) and working ($\mathbf{R}_W$) correlation structures, as listed in Table 1, where we set effect size $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ as $(1, 0.5, 0.4, 0.1)$, $\rho = 0.5$, type I error rate at 0.05, and visit times at $0, 2, 4, 6$. Under a design of complete randomization, we choose a 50:50 assignment of subjects into two treatment arms. In Figure 2.1, we see that the sample size increases as power increases and QIF sample size is smaller than (or equal to) GEE's when the working correlation is the same as (or different than) the true one.

Table 2.1: A comparison of the sample size requirements by GEE and QIF analysis with QIF-calculated sample size and percentage of reduction in the parenthesis.

		Working correlation	
True	Null	CS	AR-1
Independent	$\beta_1 = 0$	$665(642, 3.5\%)$	$690(642, 7.0\%)$
	$\beta_1 = \beta_3 = 0$	$321(281, 12.5\%)$	$285(281, 1.4\%)$
CS	$\beta_1 = 0$	$\mathbf{634}(\mathbf{634}, \mathbf{0}\%)$	$702(635, 9.5\%)$
	$\beta_1 = \beta_3 = 0$	$\mathbf{643}(\mathbf{643}, \mathbf{0}\%)$	$682(644, 5.6\%)$
AR-1	$\beta_1 = 0$	$811(739, 8.9\%)$	$\mathbf{736}(\mathbf{736}, \mathbf{0}\%)$
	$\beta_1 = \beta_3 = 0$	$603(565, 6.3\%)$	$\mathbf{565}(\mathbf{565}, \mathbf{0}\%)$
1-dep	$\beta_1 = 0$	$899(773, 14.0\%)$	$757(749, 1.1\%)$
	$\beta_1 = \beta_3 = 0$	$557(479, 14.0\%)$	$483(479, 0.8\%)$
UN	$\beta_1 = 0$	$605(600, 0.8\%)$	$693(616, 11.1\%))$
	$\beta_1 = \beta_3 = 0$	$527(520, 1.3\%)$	$601(535, 11.0\%)$

Example 2: We fix the power to be 80% and use the following matrix true

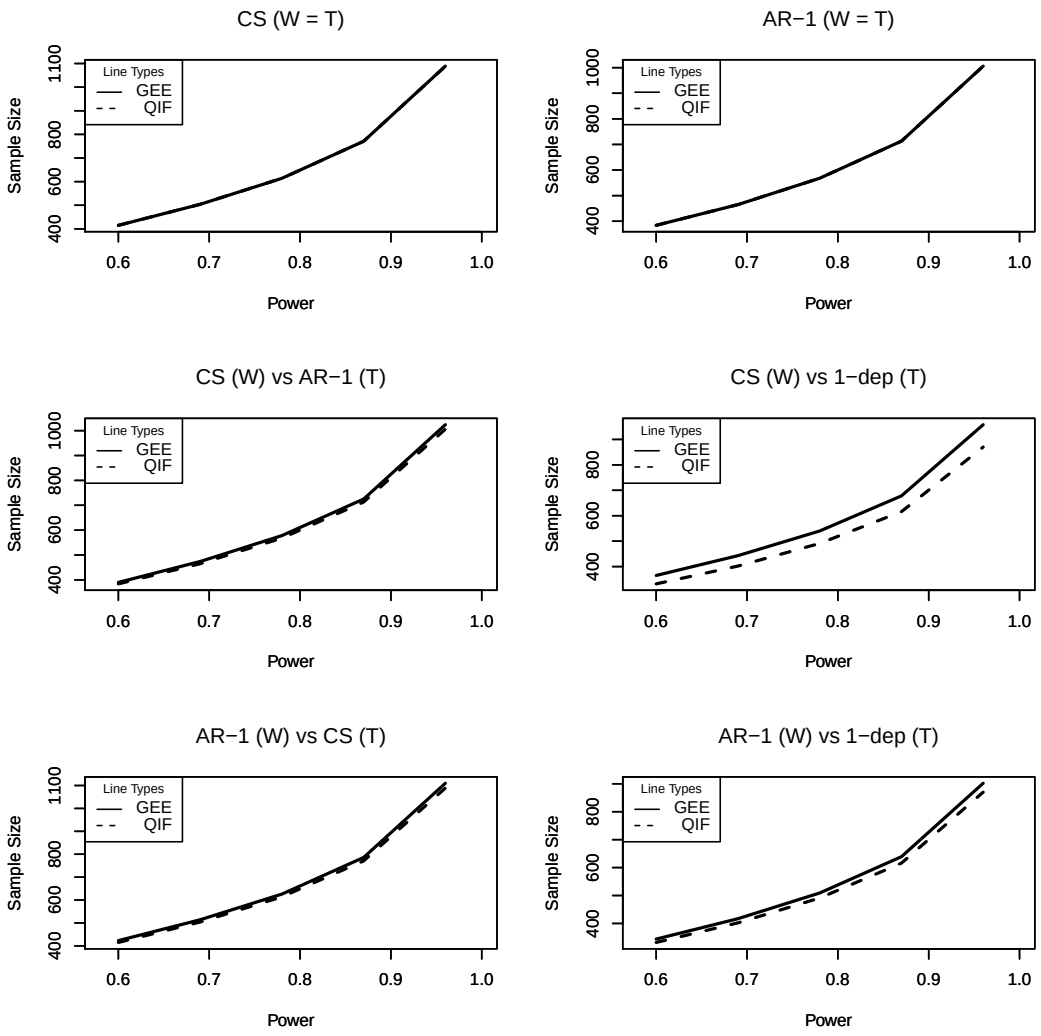

Figure 2.1: Sample size comparison for testing joint treatment effect between QIF and GEE by power. Effect size: $(1, 0.5, 0.4, 0.1)$, correlation: 0.5, type I error rate: 0.05, time points: $0, 2, 4, 6$.

unstructured (UN) correlation matrix:

$$\mathbf{R}_T = \begin{pmatrix} 1 & 0.4 & 0.3 & 0.2 \\ & 1 & 0.5 & 0.4 \\ & & 1 & 0.6 \\ & & & 1 \end{pmatrix}. \tag{2.10}$$

When the correlation matrix is correctly specified ($\mathbf{R}_T = \mathbf{R}_W$), the GEE and QIF analysis require the same sample size, as indicated by the bolded numbers in Table 1. The 1-dependence structure means that only the measurements at two nearby time points are correlated. Mathematically, it is characterized by a width-3-banded matrix - with elements 1's along the main diagonal and ρ at the first upper and lower diagonals. This table also clearly indicates the sample size saving of the QIF analysis in all cases with misspecified correlation structures, and the amount of saving by the QIF varies from 1.9% to 10.1%, depending on how severely the working correlation matrix deviates from the true correlation structure. Clearly, the QIF-based design is advantageous over the GEE-based design in term of study cost saving, when the working correlation structure used in the design is different from the true one. Misspecification is indeed the case in practice.

Example 3: Arguably, knowing the true correlation matrix of the data that are not yet collected is not possible. Figure 2.2 displays several examples in which the minimal average risk strategy is applied to the optimal sample size. All the panels on the left column are the sample size distributions when the working CS structure is used in the design, while those on the right column are based on the use of the working AR-1 structure in the design. As seen, in all the cases the mean

sample size from the QIF study design appears consistently smaller than that of the GEE study design. The top two panels are the sample size distributions when the matrix prameter $\mathbf{R}_0$ in the Wishart distribution is specified as the chosen working correlation. The middle and bottom panels are based on unstructured $\mathbf{R}_0$ given in (2.10) and 1-dependence, respectively. Clearly, at the same type I error and power level, the QIF sample size is on average smaller than the GEE sample size, even if the correlation matrices are sampled from a Wishart distribution centered at the working correlation matrix (as in the top panels). When the correlation matrix is simulated from a Wishart distribution with the $\mathbf{R}_0$ being a 1-dependence matrix, the percentage of QIF sample size saving relative to the GEE is about 6%. When the prior correlation matrix $\mathbf{R}_0$ is unstructured, the two designs require very similar sample sizes.

To demonstrate the relation of n and m, in Figure 2.3, we set power as 0.8, type I error as 0.05 and $\rho = 0.5$. When $m = 2$, it is easy to show that the sample size for the two designs is effectively the same. An analytic explanation for this equivalency can be found in Appendix A. Again, as shown by the top two panels, the QIF sample size requirement is identical to that of GEE when no misspecification is present. With misspecification, all the other four panels indicate in Figure 2.3 that QIF requires smaller sample sizes. To be specific, the percent of sample size saving by QIF is up to 47.6%, 54.2%, 42.7% and 3.9% respectively. Also, under misspecification, the GEE analysis tends to require a larger sample size with a greater m. As illustrated in section 2.4, this is probably due to the fact that the degree of misspecification increases as m grows and thus more subjects are needed to fight against the difference between working and true correlation matrices. This phenomenon is also obvious in Figure 2.4, where we test for the

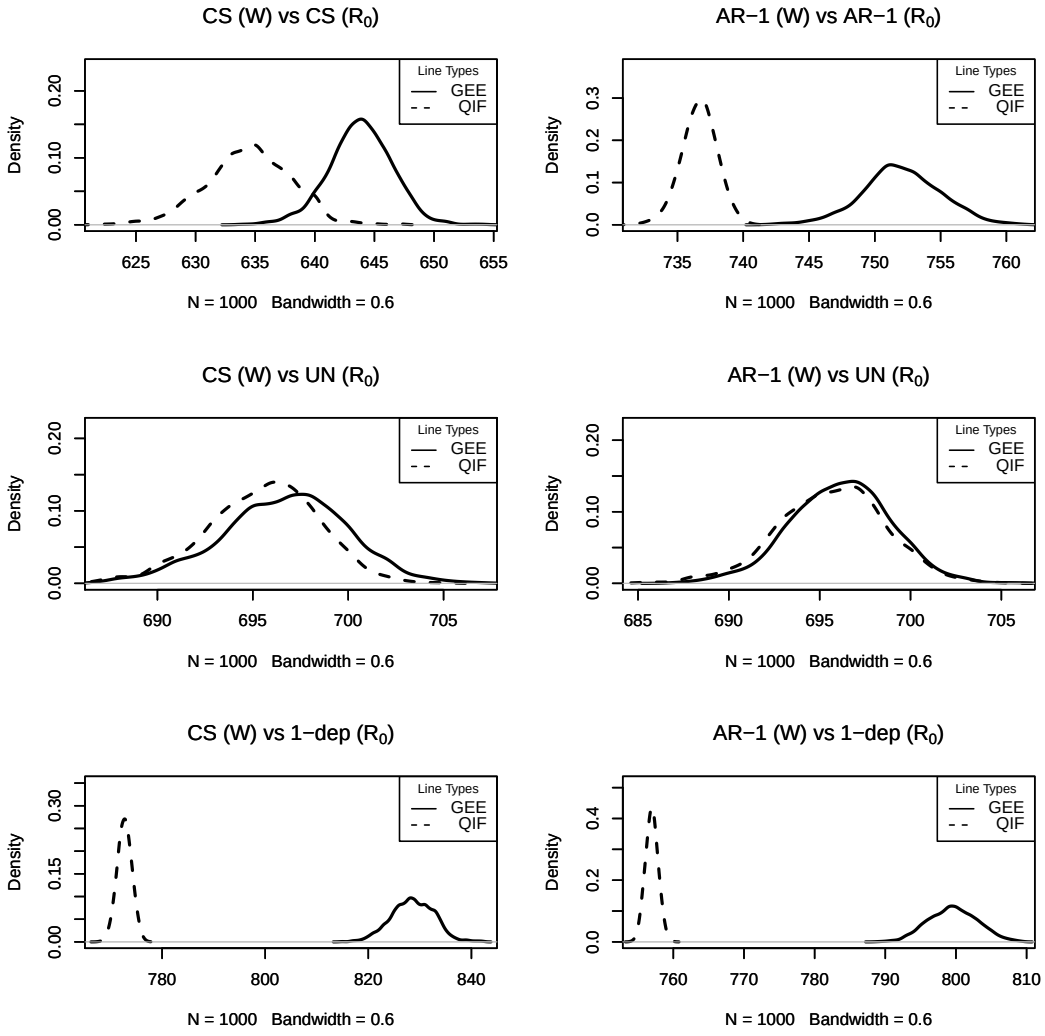

Figure 2.2: Distributions of the sample size required by the GEE and QIF analysis. R_0 denotes the prior correlation structure in the Wishart distribution. The top two panels correspond to using the working correlation as the matrix parameter for the Wishart distribution.

total and main Rx effect. We should notice that this trend is substantially reduced for the QIF method, possibly because QIF uses only basis matrices and hence is more robust to misspecification.

As requested by a reviewer, we have used the setting of Example 1 to perform a sanity check on our sample size formulas. For simplicity, we considered $m = 3$ (i.e. visits at $0, 2, 4$) and simulated 1000 sets of correlated binary outcomes under $\beta_1 = 0$ and 0.5. Based on model (2.5), we calculated empirical type I error ($\beta_1 = 0$) and power ($\beta = 0.5$) in the scenarios with and without correlation misspecification. We summarize the results for two cases of no correlation misspecification as following: (i) the true and working correlation are both CS, the power for both GEE and QIF is 0.8 and both type I error rates are 0.048; and (ii) the true and working correlation are AR-1, both power maintains as 0.8 and both type I errors are 0.052. Moreover, we summarize the results for the two cases of misspecified correlation as: (i) under a case of the true as CS and the working as AR-1, the power is 0.81 for both methods and the type I errors are 0.052 and 0.054, respectively, for GEE and QIF; and (ii) when the true is AR-1 and the working is CS, the power is 0.8 for GEE and 0.83 for QIF, while the corresponding type I errors are 0.058 for GEE and 0.053 for QIF. All these cases have confirmed the desirable control of type I error and power using the given sample sizes obtained from our formulas.

2.4 Relation to number of repeated measurements

A special operating characteristic of a longitudinal study is the number of repeated measurements (m). Under the ideal and simple scenarios (*Diggle et al.*, 2002; *Pan*, 2001a), increasing m results in smaller n. Many clinical practitioners

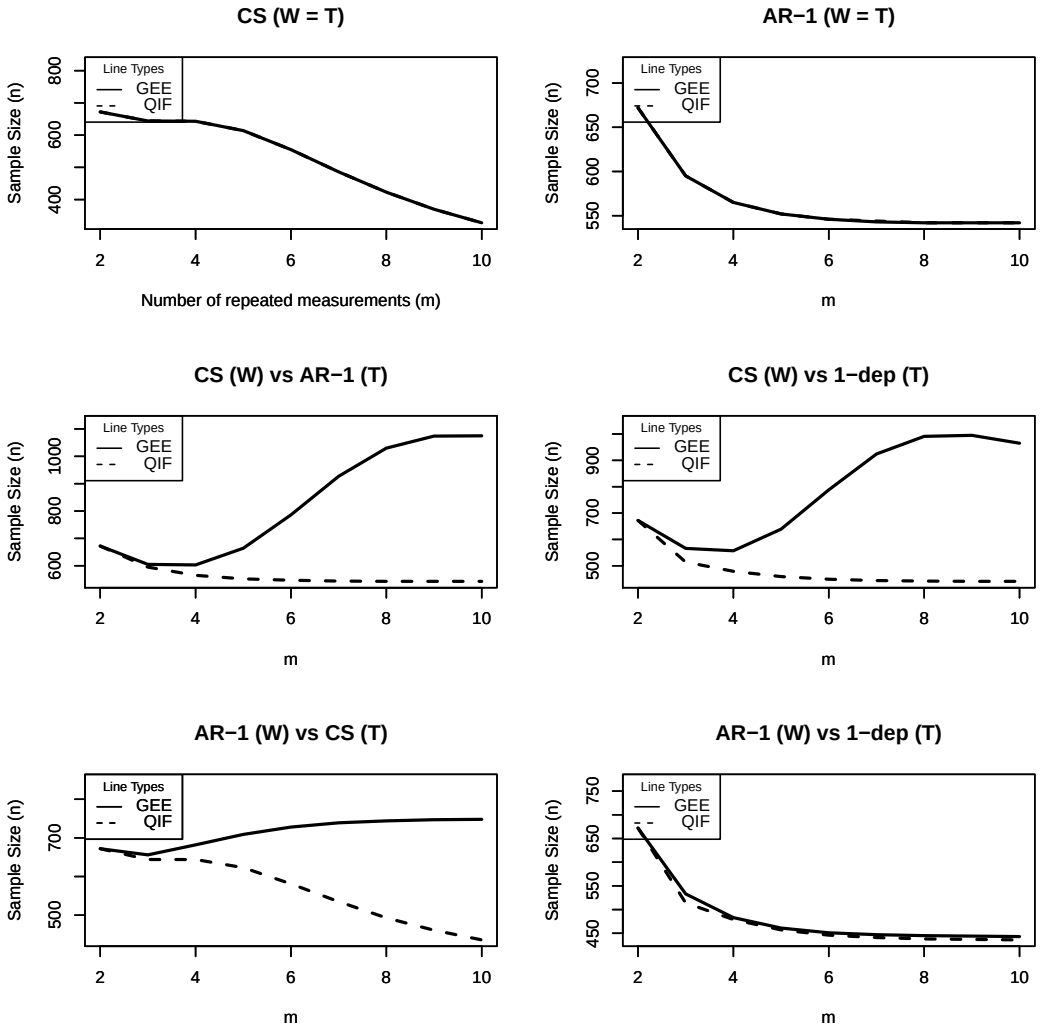

Figure 2.3: Sample size comparison between QIF and GEE for testing joint treatment effect with varying number of repeated measurements for each individual when misspecification occurs. The top panels consider cases with no misspecification; the middle panels use CS working correlation and the bottom panels take AR-1 working correlation for both designs.

take this property for granted, and use it as an effective remedy to the insufficiency in n for their studies. In fact, the relationship between n and m is far more complex than the monotonic inverse proportionality in the simple GEE settings.

When there is no misspecification, increasing m inherently produces more data, which is equivalent to increasing sample size. This means that if we can elongate the study period, we can recruit less patients. The top panels in Figure 2.2 verifies this statement. However, if there is misspecification and we are using a working correlation to approximate the true one, we then are introducing error in our modeling process. As demonstrated in *Overall and Tonidandel* (2004), the degree of misspecification would affect the power of GEE analysis. If the gained information from increasing m is not sufficient to account for the error, we certainly will need to increase our sample size to maintain the power of the study.

To facilitate the discussion, let us first look at some numerical evidence shown in Figure 2.3 where the null hypothesis is $H_0 : \beta_1 = \beta_3 = 0$ (joint treatment effect). Using the setup of Example 1 in section 3.4, we take a snapshot fixed at power 0.8, type I error 0.05 and $\rho = 0.5$. It is easy to visualize that (i) when $m = 2$, the sample size for the two designs is the same; (ii) when the working correlation is specified the same as the true correlation, as indicated by the plots in the top two panels, the QIF sample size requirement is identical to that of GEE and the sample size decreases as the number of follow-up visits increases; (iii) three of the remaining plots in the middle and bottom panels, corresponding to the mismatched working correlation structures to the true ones, clearly indicate opposite behaviors: not only does the QIF method require smaller sample sizes but it also appears rather robust to the increased number of follow-up visits. However, the GEE design is sensitive to varying numbers of follow-up visits. Figure 3 confirms the evidence

drawn from Figure 2 under the other two null hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3$ (total effect) and $H_0 : \beta_1 = 0$ (main treatment effect).

We have attempted to provide some analytic insights as how the sample size n and the cluster size m would behave. However, it is difficult to provide a general theory. In section 4.1, we can analytically prove that when the $\mathbf{R}_W$ used in the design is the same as the true $\mathbf{R}_T$, the monotonic inverse proportionality between n and m still holds. For the case of misspecification, in section 4.2 we provide analytic arguments in an example to disprove the relationship.

2.4.1 Correctly specified correlation structure

When there is no correlation misspecification, the sensitivity matrix $\mathbf{S}_n(\boldsymbol{\beta})$ is the same as the variability matrix $\mathbf{W}_n(\boldsymbol{\beta})$ for GEE. Hence, the Godambe information matrix reduces to the sensitivity matrix $\mathbf{S}_n(\boldsymbol{\beta})$, given by (2.14) in Appendix B. Further, we partition two treatment-arm specific sensitivity matrices, $\mathbf{S}_{1,n}(\boldsymbol{\beta})$ and $\mathbf{S}_{0,n}(\boldsymbol{\beta})$, into 4×4 matrices, with elements denoted as $Q_{ij}^{(1)}$ and $Q_{ij}^{(0)}, i, j = 1, \ldots, 4$.

In the following we focus on two cases where the CS and AR-1 structures are used in the design. We show that at a given power, the monotonicity relationship between n and m remains. Our argument shows that as m increases, all of $Q_{11}^{(1)}$, $Q_{33}^{(1)}$, $Q_{11}^{(0)}$ and $Q_{33}^{(0)}$ increase, and hence the resulting standard errors decrease, which leads to smaller sample sizes.

It is easy to show that for the CS working correlation,

$$Q_{11}^{(1)} = -\frac{(m-2)\rho + 1}{(\rho - 1)[(m-1)\rho + 1]} \sum_j V_{1j} + \frac{\rho}{(\rho - 1)[(m-1)\rho + 1]} \sum_{j \neq k} V_{0j}^{\frac{1}{2}} V_{1k}^{\frac{1}{2}}.$$

When $\sum_{j\neq k} V_{0j}^{\frac{1}{2}} V_{1k}^{\frac{1}{2}} = O(m)$[1], we have

$$Q_{11}^{(1)} = \frac{\rho}{1-\rho} \sum_{j=1}^{m} V_{1j} + o(1), \quad \text{as} \quad m \to \infty.$$

Since $V_{1j} > 0$, then $\sum_{j=1}^{m} V_{1j}$ increases as m increases.

Similarly, for the case of AR-1 correlation, we have

$$Q_{11}^{(1)}(m) = \frac{1+\rho^2}{1-\rho^2} \sum_{j=1}^{m} V_{1j} - \frac{2\rho}{1-\rho^2} \sum_{j=1}^{m-1} V_{1j}^{\frac{1}{2}} V_{1,j+1}^{\frac{1}{2}} - \frac{\rho^2}{1-\rho^2} V_{11} - \frac{\rho^2}{1-\rho^2} V_{1m}$$

which leads to the observation

$$Q_{11}^{(1)}(m+1) - Q_{11}^{(1)}(m) \;\; = \;\; \frac{1}{1-\rho^2} (V_{1,m+1}^{\frac{1}{2}} - \rho V_{1m}^{\frac{1}{2}})^2 \geq 0,$$

indicating that $Q_{11}^{(1)}(m)$ increases as m increases.

2.4.2 Misspecified correlation structure

When the correlation structure is misspecified, the inverse monotonicity relationship between m and n no longer holds. It is interesting to note that the magnitude of deviation between variability matrix $\mathbf{W}(\beta)$ and sensitivity matrix $\mathbf{S}(\beta)$ affects the properties of this relationship. Similar to the measurement of goodness of fit by *Pan* (2001b) and *White* (1980), we adopt the following notation d_R to quantify the deviation:

$$d_R = \|\mathbf{R}_W^{-1} \mathbf{R}_T \mathbf{W}_S^{-1} - \mathbf{R}_W^{-1}\|_2 = \|\mathbf{R}_W^{-1} (\mathbf{R}_T - \mathbf{R}_W) \mathbf{R}_W^{-1}\|_2, \tag{2.11}$$

[1] $o(x)$ means increase in a smaller rate than x.

with $\|\mathbf{A}\|_2$ denoting the L_2-norm $\|\mathbf{A}\|_2 = \max_i \sqrt{\sum_j a_{ij}^2}$.

The discrepancy measure d_R in (2.11) does not have a closed analytical form except in some simple cases, but can be easily obtained numerically. In Figure 2.3, we have seen in the middle left panel that the sample size required by GEE increases as m increases. This panel corresponds to the combination of CS as working and AR-1 as true correlation. We find similar distortion for testing total effect and the main treatment effect in Figure 2.4. To further explore this panel in terms of the relationship of sample size to the deviation d_R, we provide four additional plots in Figure 2.5 with $\rho = 0.1$ and $\rho = 0.9$, respectively. This figure indicates clearly that the inverse monotonicity relationship between n and m is distorted in the GEE analysis, but interestingly the QIF analysis is barely affected. What is also striking is that the severity of such distortion worsens for larger d_R (and ρ). We have tried other pairs of R_T and R_W scenarios, and found very similar properties. One possible explanation is that the GEE sample size is critically dependent on the chosen working correlation matrix and how severely it deviates from the true correlation matrix.

2.5 Concluding remarks

In this paper, we have developed the sample size and power calculation for both QIF and GEE analysis for dichotomous outcomes, and made detailed comparisons between these two designs. We showed that the QIF approach enjoys a sample size saving over the GEE approach; on some occasions, the saving is substantial. We anticipate that such a benefit of sample size saving remains in other types of outcome variables when correlation structure is misspecified.

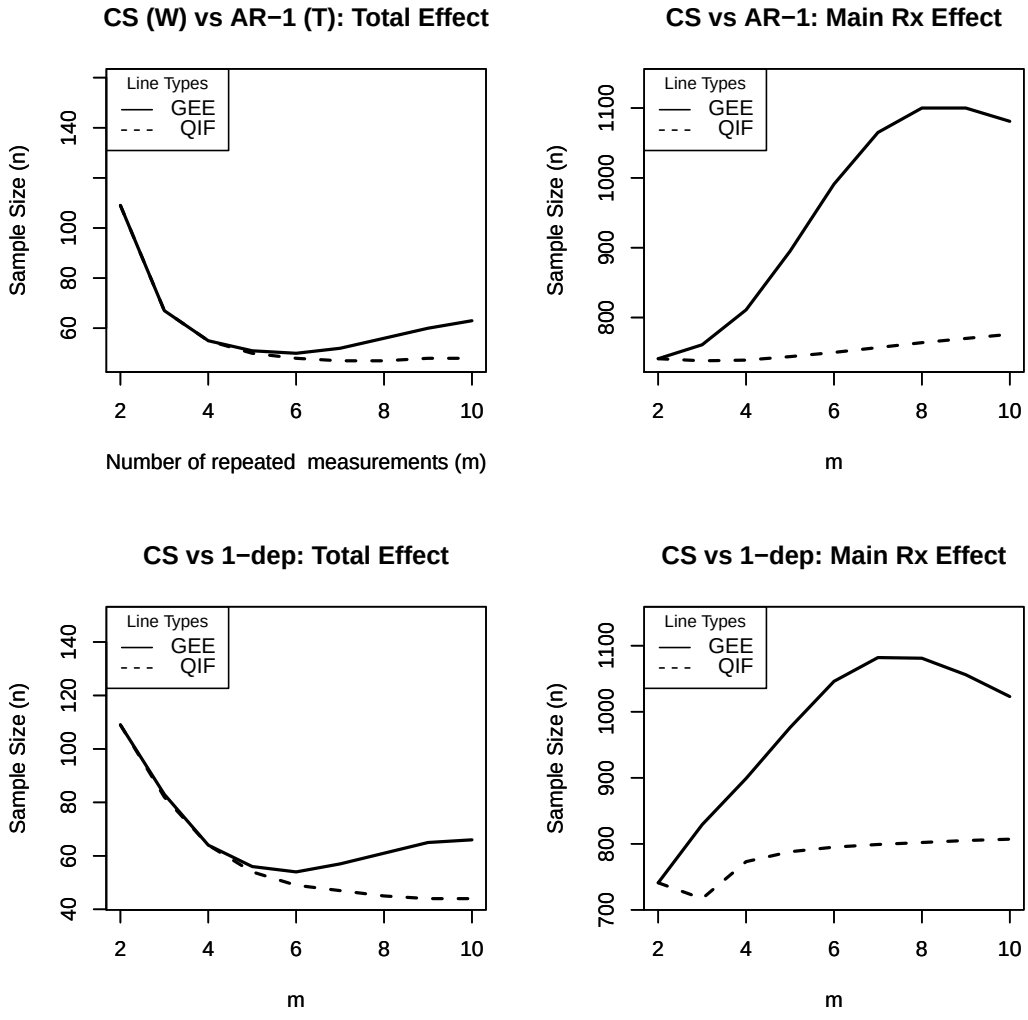

Figure 2.4: Sample size comparison between QIF and GEE for testing total effect and main Rx effect with varying number of repeated measurements for each individual when R_T is CS.

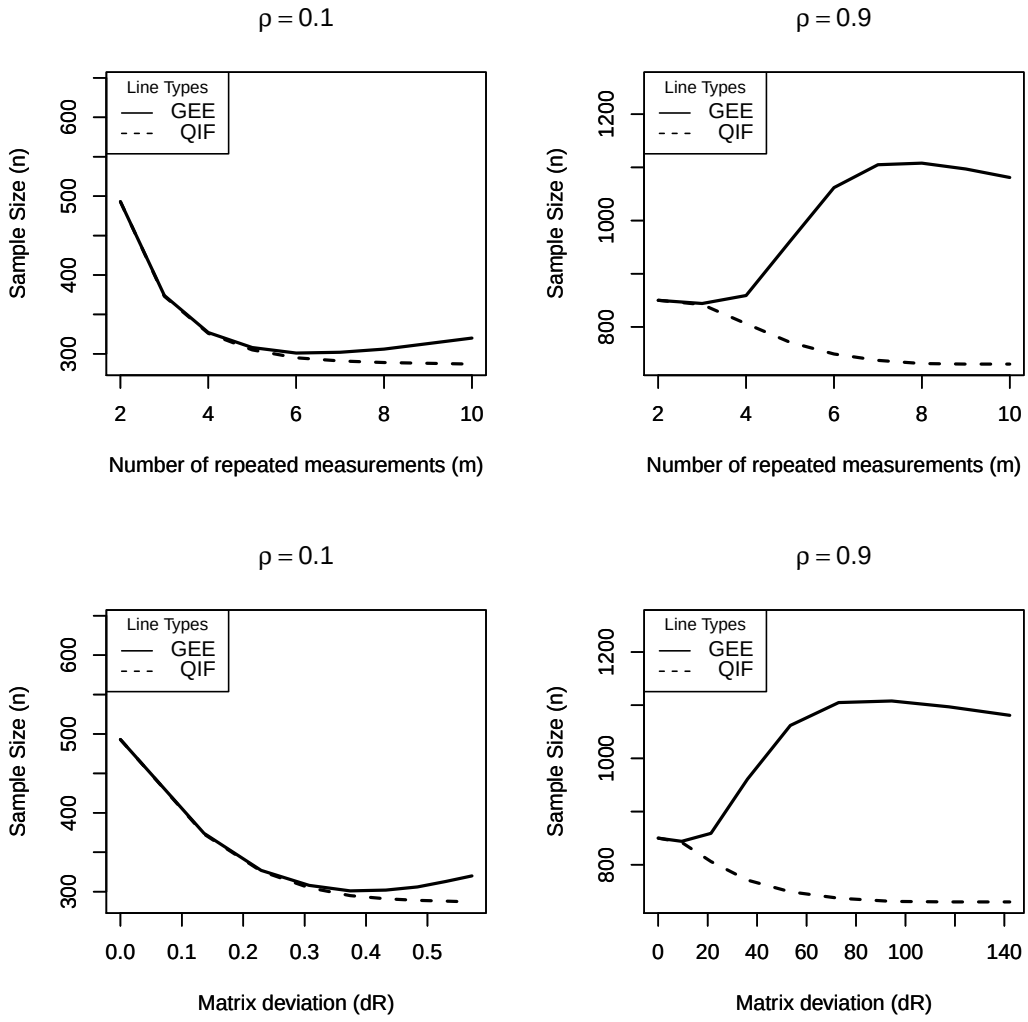

Figure 2.5: A demonstration of the trend of GEE requiring more sample sizes as the number of repeated measurements and the deviation of true and working correlation matrices increase.

We proposed an optimal sample size determination in terms of minimal average risk in the scenario where the true correlation structure is unknown. Our strategy is to vary the sample size among possible correlation structures simulated from a prior Wishart distribution, and then take the averaged sample size to be used in an actual design. We argue this strategy is optimal in terms of minimal average risk.

In addition, we demonstrate the robust behavior of the QIF sample size in response to an increased number of follow-up visits, in contrast to GEE, which requires more subjects in order to follow patients over more visits. We regard this as an important property and a clear advantage of the QIF analysis over the GEE analysis as it can reduces the burden of subject recruitment and hence cost of study.

We detail an R package used to determine GEE and QIF sample sizes is detailed in appendix D and make it available from `http://www-personal.umich.edu/~pxsong`.

2.6 Appendix

In appendix A we provide the details concerning the $\mathbf{C}_n(\beta)$ matrix. Appendix B presents the detail of the GEE sample size calculation for the logistic model (2.5), and appendix C includes some examples of the R package QIFSAMS.

A. $\mathbf{C}_n$ MATRIX IN QIF

It is easy to see that exchanges between rows or column on $\overline{\mathbf{g}}_n$ will not change the matrix $\mathbf{C}_n(\boldsymbol{\beta})$. We use symbol $\breve{\mathbf{X}}$ to denote the post-exchange variable $\mathbf{X}$ and

hence have

$$\breve{\mathbf{g}}_i' = (\mathbf{1}', \mathbf{t}_i', \boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}, \tilde{\boldsymbol{\omega}}_i'\mathbf{A}_i^{-\frac{1}{2}}, d_i\mathbf{1}, d_i\mathbf{t}_i', d_i\boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}, d_i\tilde{\boldsymbol{\omega}}_i'\mathbf{A}_i^{-\frac{1}{2}})'\mathbf{e}_i \stackrel{def}{=} (\mathbf{Z}_i, d_i\mathbf{Z}_i)'\mathbf{e}_i \stackrel{def}{=} \breve{\mathbf{B}}_i\mathbf{e}_i, \tag{2.12}$$

where $\mathbf{Z}_i = (\mathbf{1}', \mathbf{t}_i', \boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}, \tilde{\boldsymbol{\omega}}_i'\mathbf{A}_i^{-\frac{1}{2}})'$. We then have

$$\begin{aligned}\breve{\mathbf{C}}(\boldsymbol{\beta}) &= \bar{d}\breve{\mathbf{B}}_1\mathbf{A}_1^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_1^{\frac{1}{2}}\breve{\mathbf{B}}_1 + (1-\bar{d})\breve{\mathbf{B}}_0\mathbf{A}_0^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_0^{\frac{1}{2}}\breve{\mathbf{B}}_0 \\ &= \begin{pmatrix}\mathbf{G} & \mathbf{F} \\ \mathbf{F} & \mathbf{F}\end{pmatrix},\end{aligned}$$

where

$$\begin{aligned}\mathbf{G} &= \bar{d}\mathbf{Z}_1\mathbf{A}_1^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_1^{\frac{1}{2}}\mathbf{Z}_1' + (1-\bar{d})\mathbf{Z}_0\mathbf{A}_0^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_0^{\frac{1}{2}}\mathbf{Z}_0', \\ \mathbf{F} &= \bar{d}\mathbf{Z}_1\mathbf{A}_1^{\frac{1}{2}}\mathbf{R}_T\mathbf{A}_1^{\frac{1}{2}}\mathbf{Z}_1'.\end{aligned}$$

For $m \geq 2$, matrix $\mathbf{F}$ is singular since the 4-th row in $\mathbf{Z}_i$ is a linear combination of the other three. This leads to the singularity of matrix $\breve{\mathbf{C}}_n(\boldsymbol{\beta})$, which can be overcome by removing the 4-th row. Then an inference function is given by

$$\hat{\mathbf{g}}_i = (\mathbf{1}'\mathbf{e}_i, \mathbf{t}_i'\mathbf{e}_i, \boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}\mathbf{e}_i, d_i\mathbf{1}e_i, d_i\mathbf{t}_i'\mathbf{e}_i, d_i\boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}\mathbf{e}_i)',$$

which was used throughout this paper under the CS correlation when $m \geq 3$. We also removed $\boldsymbol{\omega}_i'\mathbf{A}_i^{-\frac{1}{2}}$ when $m = 2$. It should be aware that for this case the inference functions of QIF and GEE are the same and hence their calculated sample sizes are the same.

The same procedure was taken for the case of AR-1 working correlation.

B. GODAMBE INFORMATION MATRIX FOR GEE

To obtain the GEE Godambe information matrix for the logistic model (2.5), we find that the sensitivity matrix is

$$\begin{aligned} \mathbf{S}_n(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{A}_i \Sigma_i^{-1} \mathbf{A}_i \mathbf{D}_i \qquad (2.13) \\ &= -\bar{d} n \mathbf{X}_1 \mathbf{A}_1^{\frac{1}{2}} \mathbf{R}_W^{-1} \mathbf{A}_1^{\frac{1}{2}} \mathbf{X}_1' - (1-\bar{d}) n \mathbf{X}_0 \mathbf{A}_0^{\frac{1}{2}} \mathbf{R}_W^{-1} \mathbf{A}_0^{\frac{1}{2}} \mathbf{X}_0'. \qquad (2.14) \end{aligned}$$

The variability matrix is

$$\begin{aligned} \mathbf{W}_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{A}_i \Sigma_i^{-1} Var(\mathbf{r}_i) \Sigma_i^{-1} \mathbf{A}_i \mathbf{D}_i \\ &= \bar{d} n \mathbf{X}_1 \mathbf{A}_1^{\frac{1}{2}} \mathbf{R}_W^{-1} \mathbf{R}_T \mathbf{R}_W^{-1} \mathbf{A}_1^{\frac{1}{2}} \mathbf{X}_1' + (1-\bar{d}) n \mathbf{X}_0 \mathbf{A}_0^{\frac{1}{2}} \mathbf{R}_W^{-1} \mathbf{R}_T \mathbf{R}_W^{-1} \mathbf{A}_0^{\frac{1}{2}} \mathbf{X}_0'. \end{aligned}$$

Then, the Godambe information matrix is calculated by $\mathbf{J}_n(\boldsymbol{\beta}) = \mathbf{S}_n(\boldsymbol{\beta})' \mathbf{W}_n(\boldsymbol{\beta})^{-1} \mathbf{S}_n(\boldsymbol{\beta})$. The above sensitivity can be partitioned in the following fashion. When there is no model misspecification, denoting $R_W^{-1} = (R_{jk})_{mm}$, we find

$$\mathbf{S}_{1,n}(\boldsymbol{\beta}) = \begin{pmatrix} Q_{11}^{(1)} & Q_{11}^{(1)} & Q_{13}^{(1)} & Q_{13}^{(1)} \\ Q_{11}^{(1)} & Q_{11}^{(1)} & Q_{13}^{(1)} & Q_{13}^{(1)} \\ Q_{13}^{(1)} & Q_{13}^{(1)} & Q_{33}^{(1)} & Q_{33}^{(1)} \\ Q_{13}^{(1)} & Q_{13}^{(1)} & Q_{33}^{(1)} & Q_{33}^{(1)} \end{pmatrix}, \quad \mathbf{S}_{0,n}(\boldsymbol{\beta}) = \begin{pmatrix} Q_{11}^{(0)} & 0 & Q_{13}^{(0)} & 0 \\ 0 & 0 & 0 & 0 \\ Q_{13}^{(0)} & 0 & 0 & 0 \\ 0 & 0 & 0 & Q_{33}^{(0)} \end{pmatrix},$$

where

$$
\begin{aligned}
Q_{11}^{(1)} &= \sum_{j,k} R_{jk} V_{1j}^{\frac{1}{2}} V_{1k}^{\frac{1}{2}}, \qquad Q_{11}^{(0)} = \sum_{j,k} R_{jk} V_{0j}^{\frac{1}{2}} V_{0k}^{\frac{1}{2}} \\
Q_{13}^{(1)} &= \sum_{j,k} R_{jk} t_k V_{1j}^{\frac{1}{2}} V_{1k}^{\frac{1}{2}}, \qquad Q_{13}^{(0)} = \sum_{j,k} R_{jk} t_k V_{0j}^{\frac{1}{2}} V_{0k}^{\frac{1}{2}} \\
Q_{33}^{(1)} &= \sum_{j,k} R_{jk} t_j V_{1j}^{\frac{1}{2}} t_k V_{1k}^{\frac{1}{2}}, \qquad Q_{33}^{(0)} = \sum_{j,k} R_{jk} t_j V_{0j}^{\frac{1}{2}} t_k V_{0k}^{\frac{1}{2}}.
\end{aligned}
$$

C. R Package: QIFSAMS

The R package QIFSAMS includes two functions that calculate sample size for the QIF and GEE analysis, respectively. This package and the user's manual can be downloaded from wepage: http://www.umich.edu/∼pxsong.

To obtain the sample size, the user needs to specify effect sizes and correlation matrices. QIFSAMS takes effect sizes of zero time and/or interaction effect as well. QIFSAMS requires $m \geq 2$ to reflect a longitudinal study design. For example, the following commands were used to produce Table 1.

```
GEE.n <- GEE_Size(typeIerror = 0.05, power = 0.8, coeff = c(1,0.5,0.4,0.1),
 num_repeated = 4, ratioTRT = 0.5,timeUnit = 2, visits = c(0,2,4),
 corr_T = 0.5, corr_W = 0.5, R_wk_ID = 1, R_true_ID = 1,R_un =
 corr.matrix.un, test.ID = 1)
QIF.n <- QIF_Size(typeIerror = 0.05, power = 0.8, coeff = c(1,0.5,0.4,0.1),
 num_repeated = 4, ratioTRT = 0.5,timeUnit = 2, visits = c(0,2,4),
 corr = 0.5 ,R_wk_ID = 1, R_true_ID = 1, R_un = corr.matrix.un, test.ID = 1).
```

Arguments Type I error, power, effect size, number of repeated measurements, proportion of subjects in the test treatment group, time points, correlation parameters need to be specified in the functions. Note that in GEE, we need to

specify both true and working correlation parameters but in QIF we only need the true correlation parameters. For example, when both true and working correlation are CS, as shown above, we set both `R_wk_ID` and `R_true_ID` to be 1 and obtain the corresponding two sample sizes. For other cases in Table 1, we vary these two parameters.

Both functions combines `R_true_ID = 4` and `R_un` to specify a UN correlation, which can be sampled from a Wischart distribution. In Figure 2.2, we generated 1000 such samples, choose a working correlation and then obtain the sample size distributions.

More details can be found from the help file from the package.

CHAPTER III

Accurate Local Ancestry Inference in Exome Sequenced Samples

3.1 Abstract

Estimates of the ancestry of specific chromosomal regions in admixed individuals are useful for studies of human evolutionary history and in disease gene mapping. Previously, this ancestry deconvolution relied on genome-wide genotypes from high-density GWAS arrays, which are not always available in exome sequenced samples. We show that even a relatively small number of off-target reads generated during exome sequencing experiments can be used to accurately estimate the ancestry of admixed individuals. To reconstruct local ancestry, our SEQMIX method models sequence data directly, without requiring intermediate genotype calls. We evaluate the accuracy of our method through simulations and the analysis of sequenced samples in the 1,000 Genomes Project and the Women's Health Initiative. In African-Americans, we show local ancestry estimates derived using our method are extremely similar to those derived using Illumina's Omni 2.5M genotyping array (concordance between the two estimates ~ 0.97); and are

much improved compared to estimates that use only exome genotypes and ignore off-target sequencing reads. SEQMIX should be useful to anyone undertaking exome or targeted sequencing of admixed populations, both for analysis of human population history and for disease gene mapping studies.

3.2 Introduction

The genomes of admixed individuals can be described as mosaics with alternating segments of different ancestries. The pattern of each mosaic such as the origin and the length of individual segments, reflects the admixture history of each individual. Although the precise boundaries and ancestry of these segments are usually unknown, they can be reconstructed using statistical methods that examine the distribution of genetic markers along the chromosome. These methods take advantage of the differences in allele and haplotype frequencies between distinct ancestral populations.

Reconstructions of local ancestry have many uses in population genetics and in human disease gene mapping. For example, reconstructions of local ancestry have been used to investigate the genetic relationship between the admixed populations and putative ancestral groups in studies of the history of Latinos and Hispanics in North America (*Bryc et al.* 2010b) and of the Uyghur in China (*Xu et al.* 2008). Local ancestry estimates are also useful in human disease gene mapping, where they have been used to study multiple sclerosis (*Reich et al.* 2005), hypertension (*Zhu et al.* 2005), prostate cancer (*Freedman et al.* 2006), among many other diseases (see review *Winkler et al.* (2010)). In exome sequencing studies, admixture mapping can be used to improve the matching of case and control data (for exam-

ple, by stratifying comparisons between case and control chromosomes according to local ancestry).

The first applications of ancestry deconvolution relied on ancestry informative markers (AIMs) (*Smith et al.* 2004), which are carefully selected markers that show large differences in allele frequency between populations. Statistical methods used in these early applications typically use Hidden Markov Models (HMM) (*Rabiner*, 1986) and assume accurate genotypes for every marker (*Patterson* 2004). More recent methods typically do not rely on availability of AIMs but, instead, use the large amounts of data generated by GWAS arrays (where each marker provides a modest amount of information about ancestry, on average), to model local ancestry. These newer methods can still rely on Hidden Markov Models, often enhanced to model haplotype differences in addition to allele frequency differences between populations (*Price et al.* 2009,*Sundquist et al.* (2008),*Tang et al.* (2006)), but also use other statistical techniques (see review (*Seldin et al.* 2011)) such as clustering algorithms (*Sankararaman et al.* 2008) and principal component analyses (*Bryc et al.* 2010a).

Instead of GWAS arrays, the next phase of data generation for genetic studies is likely to rely on short read sequencing technologies. Targeted sequencing approaches, such as exome sequencing (*Ng et al.* 2009), are becoming increasingly popular for disease gene mapping (*Bamshad et al.* 2011) and clinical diagnosis (*Choi et al.* 2009). In exome-sequenced samples, genotypes for AIMs or GWAS SNP panels are typically not available and confident calls from targeted sequencing cover only a small portion of the genome. This poses the challenge for accurate ancestry inference.

In this paper, we show that even a relatively small number of off-target reads,

generated as a by-product of targeted sequencing experiments, can be used to accurately reconstruct the ancestry mosaic of admixed individuals. Using our method SEQMIX on simulated data, we show that – for African Americans - $> 98\%$ accurate ancestry calls can be generated with as little as 0.3-fold average off target read coverage. We also validate our approach empirically, by comparing our results with those obtained with HAPMIX in two sets of African American samples for which both GWAS array genotypes and exome sequence data are available. For both datasets with off target mean coverage at < 2x, we witness that 97% of the SEQMIX called ancestries are the same as those from running HAPMIX on genotype array genotypes for the same individuals.

SEQMIX should be useful to anyone undertaking exome or targeted sequencing of admixed populations, both for analysis of human population history and for disease gene mapping studies.

3.3 Material and Methods

3.3.1 Hidden Markov Model for Sequence Data at Unlinked Sites

Our method SEQMIX is a Hidden Markov Model (HMM) that uses exome data to infer local ancestry. Figure 1 demonstrates the three layers of the model. At each site, the bottom layer is the hidden ancestry state, the middle layer is the unknown genotype and the top layer is the observed set of sequence reads. At the middle layer, the genotype is useful in relating sequence reads and ancestry. For simplicity, we assume that all variants are biallelic and that there are two possible ancestry states. Our model can be naturally extended to accommodate multi-allelic markers or additional ancestry states.

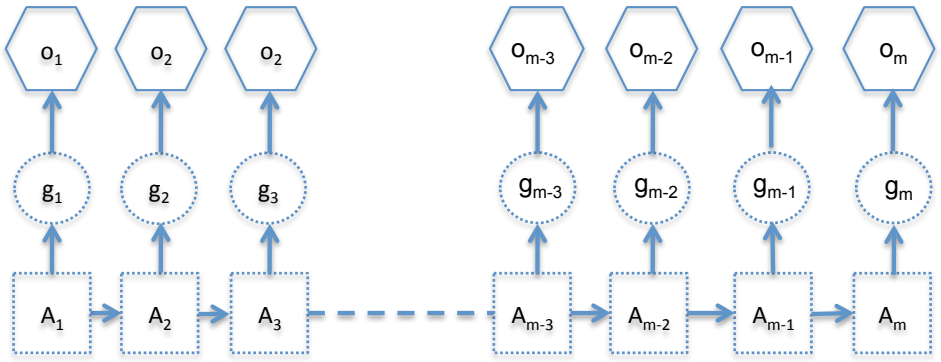

Figure 3.1: Graphic illustration of the hidden Markov model of ancestry deconvolution, assuming linkage disequilibrium bwteen markers. There are three layers of information: the square boxes (A) represent the hidden ancestry at each site, the circles (g) represent hidden genotypes and the hexagons (o) represent observed base in the sequence reads.

A Hidden Markov models has two important components: the transition matrix and emission probability. The transition probabilities for this HMM describe the probability of changes in ancestry along the chromosome. We denote the unobserved ancestry state q_s as (A_{s1}, A_{s2}), where A_{s1} indicates the ancestry of the first chromosome at site s and A_{s2} denotes the ancestry of the other chromosome. We let $\gamma_{s,s+1}$ denote the recombination rate between site s and $s+1$ and let T denote the timing of the first admixture event (measured in generations from the present). The transition matrix summarizing the probability of changes in ancestry between sites s to $s+1$ along a chromosome is then:

$$p_{s,s+1} = \begin{bmatrix} p^{E,E}_{s,s+1} & p^{E,A}_{s,s+1} \\ p^{A,E}_{s,s+1} & p^{A,A}_{s,s+1} \end{bmatrix} = \begin{bmatrix} 1-(1-e^{-\gamma_{s,s+1}T})\pi_A & (1-e^{-\gamma_{s,s+1}T})\pi_A \\ (1-e^{-\gamma_{s,s+1}T})\pi_E & 1-(1-e^{-\gamma_{s,s+1}T})\pi_E \end{bmatrix} \quad (3.1)$$

where π_A and π_E denote the prior probabilities of being an African or European ancestral allele respectively (in our analysis, $\pi_A = 0.8$ and $\pi_E = 0.2$ (*Price et al.*

2009; *Bryc et al.* 2010a)).

For a diploid individual, the transition probability must model changes along two chromosomes, and becomes $P_{s,s+1} \otimes P_{s,s+1}$, the Kronecker product of single chromosome transition matrices. Each element of this 4 by 4 matrix is defined as $t_{s,s+1}^{(i_1,j_1),(i_2,j_2)} = p_{s,s+1}^{i_1,i_2} p_{s,s+1}^{j_1,j_2}$, where $p_{s,s+1}^{i_1,i_2}$ describes the probability of changes in ancestry from state i_1 at site s to i_2 at site $s+1$ along one chromosome and $p_{s,s+1}^{j_1,j_2}$ describes the probability of changes in ancestry from state j_1 at site s to state j_2 from $s+1$.

Assuming Hardy-Weinberg Equilibrium, emission probabilities $P(o_s|q_s)$, which describe the probability of observing sequence reads o_s given an underlying ancestry state q_s, can be calculated as

$$\begin{aligned} & \mathrm{P}(o_s|q_s = (A_{s1}, A_{s2})) \\ = & \sum_{g_s} \mathrm{P}(o_s|g_s = (g_{s1}, g_{s2})) f_{A_{s1}}^{g_{s1}} (1 - f_{A_{s1}})^{1-g_{s1}} {f_{A_{s2}}}^{g_{s2}} (1 - f_{A_{s2}})^{1-g_{s2}}, \end{aligned}$$

where g_{si} is an indicator variable, which takes value 1 when the reference allele is observed in the i^{th} chromosome, and 0 otherwise. $f_{A_{si}}$ is the reference allele frequency at site s with ancestry A_{si}. The genotype likelihood $p(o_s|g_s)$ describes the probability of the observed set of reads given the underlying genotypes $g_s = g_{s1} + g_{s2}$ and is the data product of next-generation sequencing technology (*Li* 2011).

3.3.2 Linkage Disequilibrium Pruning Algorithm

The HMM described in Figure 3.1 assumes that the marker positions are in linkage equilibrium, and the observed set of sequence reads are independent given

the underlying ancestry information. But often this assumption does not hold and directly applying this model could lead to inference of many false ancestry switches. This is also a known problem for analyzing high-density genotype data (*Tang et al.* 2006).

To resolve the issue, we propose a LD pruning algorithm that attempts to balance the goals of using as much sequence information as possible and avoiding markers in linkage disequilibrium with each other. We first estimate the squared correlation ρ^2 between marker pairs within each ancestral population; for each marker pair we then record the maximum ρ^2 value observed across all populations. Then, for each individual, we list all sites with non-zero sequence coverage. For every pair of markers where ρ^2 exceeds a cutoff (we typically use 0.1), we exclude the site with lower coverage from consideration. The process continues until no marker pairs in linkage disequilibrium remain in the list of sites with non-zero coverage. Note that this pruning algorithm must be applied to each sequenced sample independently.

3.3.3 Simulation

We first evaluated our method in the analysis of simulated African American individuals. We simulated each genome using two pairs of $1,000$ Genomes Project African and European ancestry haplotypes as templates (*The 1000 Genome Project Consortium* 2010). At the beginning of each chromosome, we sampled European ancestry with probability π_E and African ancestry with probability π_A. Next, we sampled the a series of distances to the next potential ancestry switches from an exponential distribution with mean $\frac{1}{T\theta}$, where $T = 6$ is the number of generations since the first admixture event (*Price et al.* 2009) and $\theta = 10^{-8}$ (*Sankararaman*

et al. 2008) is the average recombination rate per base-pair per generation. At each switch point, we again sampled ancestry with probabilities π_E and π_A.The process continued until all chromosomes had been sampled for each individual.

Conditional on these simulated haplotypes we proceeded to generate simulated sequence data and genotype likelihoods. First, we annotate each site s as either on-target or off-target using the consensus exome capture definition from the $1,000$ Genomes Project (`http://www.1000genomes.org/data`). For on-targeted sites, we sampled the coverage C_s from a Poisson distribution with mean $\mu_{\text{on}} = 60$; for off target sites, we sampled the coverage from a Zero-inflated Poisson (ZIP) distribution (*Lambert* 2010), which is a mixture of distribution with point mass at 0 (with probability $1 - p_{\text{off}}$) and a Poisson distribution with mean μ_{off}. In our simulations, we use $\mu_{\text{off}} = 3$, which is estimated from real data. We let p_{off} vary between 0 and 0.8, so the off-targeted mean coverage varies between 0 and 2.4.

With the simulated coverage C_s and genotype g_s at site s and the assumption of a uniform sequence error rate $e = 0.01$ across all sites (*Bentley et al.* 2008), we simulated $M_s(\leq C_s)$ reads identical to reference allele using the following distribution

$$M_s|g_s = \begin{cases} \text{Binomial } (C_s, 1-e) & \text{if } \ g_s = 0 \\ \text{Binomial } (C_s, 0.5) & \text{if } \ g_s = 1 \\ \text{Binomial } (C_s, e) & \text{if } \ g_s = 2 \end{cases} . \tag{3.2}$$

The genotype $g_s = 0, 1, \text{or } 2$ is the number of copies of alternative alleles at site s. Then, we calculated the genotype likelihood at site s from the following equation

$$L(g_s) = \frac{1}{2^{C_s}}[(2-g_s)e + g_s(1-e)]^{M_s}[(2-g_s)(1-e) + g_s e]^{C_s - M_s}. \tag{3.3}$$

Finally, we ran SEQMIX on the simulated sequence data and compared the inferred ancestry to the truth. Notably, in the LD pruning step of SEQMIX, we have varied the ρ^2 cutoffs from 0.01 to 0.2 and used the physical distance times 10^{-8} to estimate the genetic distance in running SEQMIX.

3.3.4 Analysis of 49 American South West (ASW) Samples

We analyzed 49 African American samples from the 1,000 Genomes project. We ran UMAKE (Hyun M. Kang and Goo Jun in preparation) on the exome bams to obtain the genotype likelihood. We downloaded the precomputed recombination map from the IMPUTE website (*Marchini et al.* 2007) and linearly interpolated the recombination rate for sites that are not included in the map. We used $\rho^2 = 0.1$ as a cutoff for LD pruning. At last, we ran SEQMIX on the genotype likelihoods for the local ancestry results.

In addition, we ran HAPMIX on the Illumina's Omni 2.5M genotype data for these samples to infer ancestry. We used the phased European (GBR, CEU, TSI, FIN, IBS) and African (YRI, LWK) Omni genotype data as reference haplotypes and used the genetic distance map from IMPUTE. We linearly interpolated for the ancestry result at the markers in the sequence data.

Furthermore, we used both SEQMIX and HAPMIX called ancestry blocks to evaluate patterns of coding variation in proportion of the genome with recent European and African ancestry. There are three categories of ancestries at each site: both alleles are of European ancestry, both are of African ancestry and one of each kind. In each category, we calculated the heterozygosity per kilobase, and at the heterozygous sites, we counted the number of nonsynonymous sites and the number of loss-of-function sites per megabase.

3.3.5 Analysis of 79 African American Samples from Women's Health Initiative (WHI)

WHI is one of the largest studies of African American women that were undertaken in the United States. Previous description for this cohort can be found in (*The Women's Health Initiative Study Group*, 1998). We obtained exome sequence data and the GWAS genotype data for 79 samples. We calculated the genotype likelihood using UMAKE and ran SEQMIX to infer the local ancestry. Furthermore, we ran HAPMIX using the GWAS data of the 79 African American WHI samples as input and the $1,000$ Genomes Project European and African haplotypes as references. We then compared these two sets of results to evaluate the performance of SEQMIX.

3.3.6 Measuring the Performance of SEQMIX

Let $A_s^i = A_{s1}^i + A_{s2}^i$ denote the true diploid ancestry for individual i at site s, A_s^i is equal to 0 when both copies are of European ancestry and is equal to 2 when both copies are African and is equal to 1 when the diploid ancestry is one of each kind. Let X_s^i denote the estimate of A_s^i by SEQMIX.

SEQMIX uses the forward-backward algorithm to estimate the distribution of A_s^i, i.e. $P(A_s^i = j) = p_s^j, j = 0, 1, 2$. We use this value at each site along the autosomal genome to estimate the percent of whole genome European ancestry. In simulations, we compared this value calculated by the true and inferred local ancestries. For real data analysis, we compared this value calculated from results by SEQMIX and HAPMIX.

Moreover, at each variant site, we used the inferred probability p_s^j to calculate

the expected number of European copies at site s as $2p_s^0 + p_s^1$. If the number is less than 0.5, we set $X_s^i = 0$; if it is greater than 1.5, we set X_s^i as 2; otherwise, $X_s^i = 1$. To quantify the performance of SEQMIX, we calculated squared correlation as well as the concordance rate, which, for each individual, is the percent of the total m sites that have the same ancestry, calculated as $r_i = \frac{\sum_{i=1}^{m} I(X_s^i = A_s^i)}{m}$. We report the average concordance rate r across all samples. Both statistics, in analysis of the simulated data, are calculated to compare inferred and true ancestry; and in the analysis of real data, are calculated from SEQMIX and HAPMIX result.

3.4 Results

3.4.1 Simulation of Exome Sequence Data

We simulated 10 African American individuals with an average of 80% African ancestry and 20% European ancestry. We assumed the first admixture event happened 6 generations ago and used $1,000$ Genomes Project haplotypes of recent European (EUR) and African (AFR) ancestry as templates. Then we simulated the sequence reads covering each site and calculated genotype likelihoods (see Materials and Methods). Using the simulated sequence data, we ran SEQMIX to infer the local ancestry and compare the inferred result to the underlying truth.

In addition, we also investigated the impact of the ρ^2 cutoff used for pruning markers in LD with each other. In principle, low values for this cutoff might discard too many informative sites, whereas large values might result in false ancestry switches. We varied the ρ^2 cutoffs from 0.01 to 0.2 and found that for this range, there were no significant differences in terms of the concordance. Hence, we adopted 0.1 as the ρ^2 cutoff throughout this paper.

When simulating sequence coverage at each site, we sample the coverage for the on-target site from a Poisson distribution with mean 60. For the off-target sites, we sampled the coverage from a ZIP distribution with a mean varying between 0 and 2.4. Figure 3.2 illustrates accuracy of our method as a function of off-target coverage: the top panel displays the simulated true ancestry path and other panels shows the ancestry path estimated by SEQMIX with mean off-target coverage at 0, 0.003, 0.006, 0.3, 0.6, 1.2 and 2.4. As we can see from the top panel, there are 9 ancestry switches along the chromosome. When there is no off-target reads generated, 4 of the switches are missed - corresponding to the shortest ancestry tracts of length 2.1 and 4.3 Mb, respectively (second panel). If the off-target coverage is very low - for example, if only 3 sites in every thousand are covered (third panel) - the missing tracts are still not captured. But, as off target coverage increases, the missing switches are quickly recovered. When off-target mean depth reaches 0.3, all ancestry tracts and switches are detected. As off target coverage increases further, placement of the ancestry switches improves and the concordance rate approaches 1.

To better understand the effect of sequencing depth on ancestry estimates, we simulated another 100 sets of sequence data, again varying off-target coverage between 0 and 2.4. The results are summarized in Table (3.1), where we can see that the 2.1Mb European ancestry block is missed for very low off target coverage (< 0.01x) but progressively better captured as coverage increases. When coverage reaches 0.6x, this ancestry block is captured 75% of the time and the success rate reaches 93% and 95% when the coverage is 1.2x and 2.4x, respectively. For the 4.3Mb European ancestry block, 0.15x off target coverage is sufficient to achieve a 100% capture rate. The exact performance of ancestry estimation will depend

not only on block length and sequencing depth, but also on the amount of LD in a region and the fraction of each region selected for deep, targeted re-sequencing.

Table 3.2 summarizes the concordance rate across the simulated 10 samples. We can see that the average concordance and squared correlation between true and estimated ancestry increases with off-target coverage. When there are no off-target reads, the squared correlation between inferred ancestries and true ancestry is only 0.70 and ancestry estimates for 7.2% of sites are incorrect. When the off target coverage increases to 0.3x, the squared correlation reaches 0.93 and the number of sites with wrong ancestry calls drops to $\sim 2\%$. The statistics improve further as the off target coverage increases to 2.4.

Table 3.1: Percent of time that the smallest ancestry blocks are captured in the 100 sets of simulations for on-target coverage between 0 to 2.4.

	Percent of the time that the ancestry block is captured	
Mean off-target	Block size 2.1 Mb	Block size 4.3 Mb
0	0	47
0.003	0	62
0.006	0	76
0.009	0	81
0.15	39	100
0.3	69	100
0.6	75	100
1.2	93	100
2.4	95	100

3.4.2 Analysis of 49 ASW Samples

We applied SEQMIX on 49 ASW African American samples that were exome sequenced by the 1,000 Genomes Project. The average depth at on targeted sites

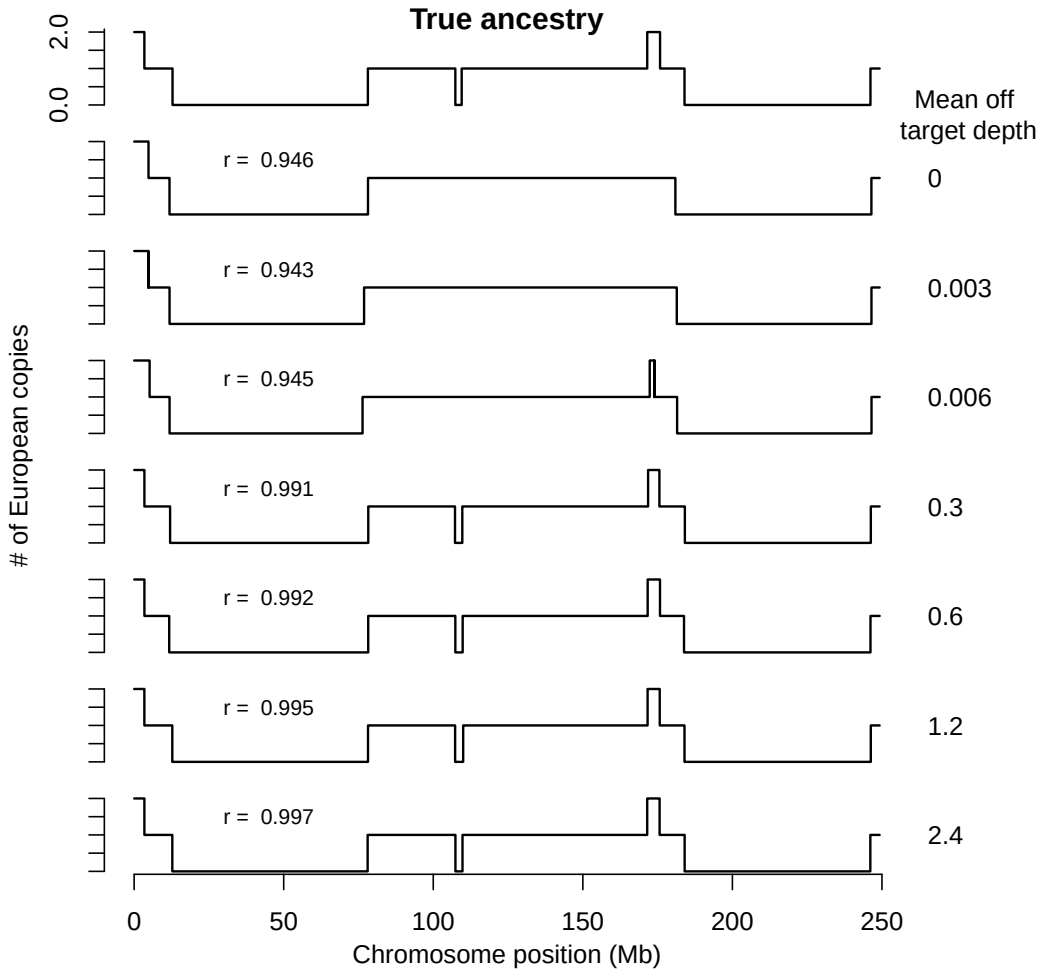

Figure 3.2: The ancestry path along chromosome 1 of a simulated African Americans with mean off target coverage varying between 0 to 2.4. For each mean off-target depth, the concordance (r) is included in the left side of each panel and the mean off-targets coverage is listed at the right side of each panel.

Table 3.2: Accuracy of SEQMIX inferred ancestry for 10 simulated admixed individuals and its relationship with the mean off targeted coverage.

Mean off-target coverage	Correlation square mean (sd)	Concordance rate mean (sd)
0	0.697 (0.181)	0.928 (0.028)
0.003	0.735 (0.151)	0.937 (0.023)
0.006	0.744 (0.183)	0.944 (0.021)
0.009	0.781 (0.150)	0.950 (0.022)
0.012	0.785 (0.144)	0.950 (0.024)
0.015	0.796 (0.122)	0.951 (0.019)
0.03	0.840 (0.093)	0.962 (0.015)
0.15	0.890 (0.070)	0.977 (0.011)
0.3	0.931 (0.036)	0.983 (0.011)
0.6	0.944 (0.040)	0.987 (0.008)
1.2	0.949 (0.040)	0.989 (0.007)
1.8	0.949 (0.042)	0.989 (0.008)
2.4	0.952 (0.035)	0.990 (0.007)

is 80 and at off target sites is 1.9. We compared ancestry estimates generated by SEQMIX (using exome sequence data as input) to estimates generated by HAPMIX (using OMNI 2.5M genotypes as input).

Among the 49 samples, one individual was estimated to have 99.2% European genome by SEQMIX, and 96.0% European ancestry by HAPMIX. Excluding this outlier, the remaining 48 samples were estimated by SEMIX to have 2.0% to 57.1% European ancestry; and by HAPMIX have 2.6% to 55.8% European ancestry. The estimates for the percent of European ancestry for coding regions are similar to those for the whole genome. The genome-wide estimates of European ancestry fraction generated by both methods are extremely similar, with squared Pearson correlation $> 99.9\%$.

Furthermore, we calculated the concordance rate and squared correlation for

the local ancestry estimates. In the top panel of Figure 3.3, we show that for the 49 samples, on average, 97% of the whole genome ancestry calls from SEQMIX and HAPMIX are the same. The concordance rate for all samples varies between 0.94 and 0.98. For the outlier whose percent European genome was reported by SEQMIX to be 99.2% but by HAPMIX to be 96.0%, the correlation square between the HAPMIX and SEQMIX called ancestries has a median of 0.89, a mean of 0.88 and standard error 0.08. The first and third quartiles for the correlation squares are 0.87 and 0.92 respectively. These numbers indicate that SEQMIX can use off target reads to infer similar results to that from using high density genotype arrays.

Another strategy to evaluate the performance of SEQMIX is to compare the estimates of some population genetic parameters using ancestry call results from SEQMIX and HAPMIX (Table 3.3). For chromosomal segments that are called both of European ancestry by SEQMIX, we calculated the heterozygosity to be 0.42 per kb with a standard error 0.09. This number is 0.43 based on HAPMIX-called-European chromosomal regions. For SEQMIX European ancestry stretches, there are 191 nonsynonymous heterozygous sites per Mb and 2.2 loss-of-function heterozygous sites per Mb. The corresponding estimates are 191 and 2.1 for HAPMIX European ancestry stretches. The equivalence of these population parameter estimates also supports that SEQMIX can effective use off-target sequence data for accurate ancestry inference.

3.4.3 Analysis of 79 WHI Samples

In addition to the 49 African American individuals, we also ran SEQMIX on a set of 79 African Americans from WHI cohort. The exome sequence data for these

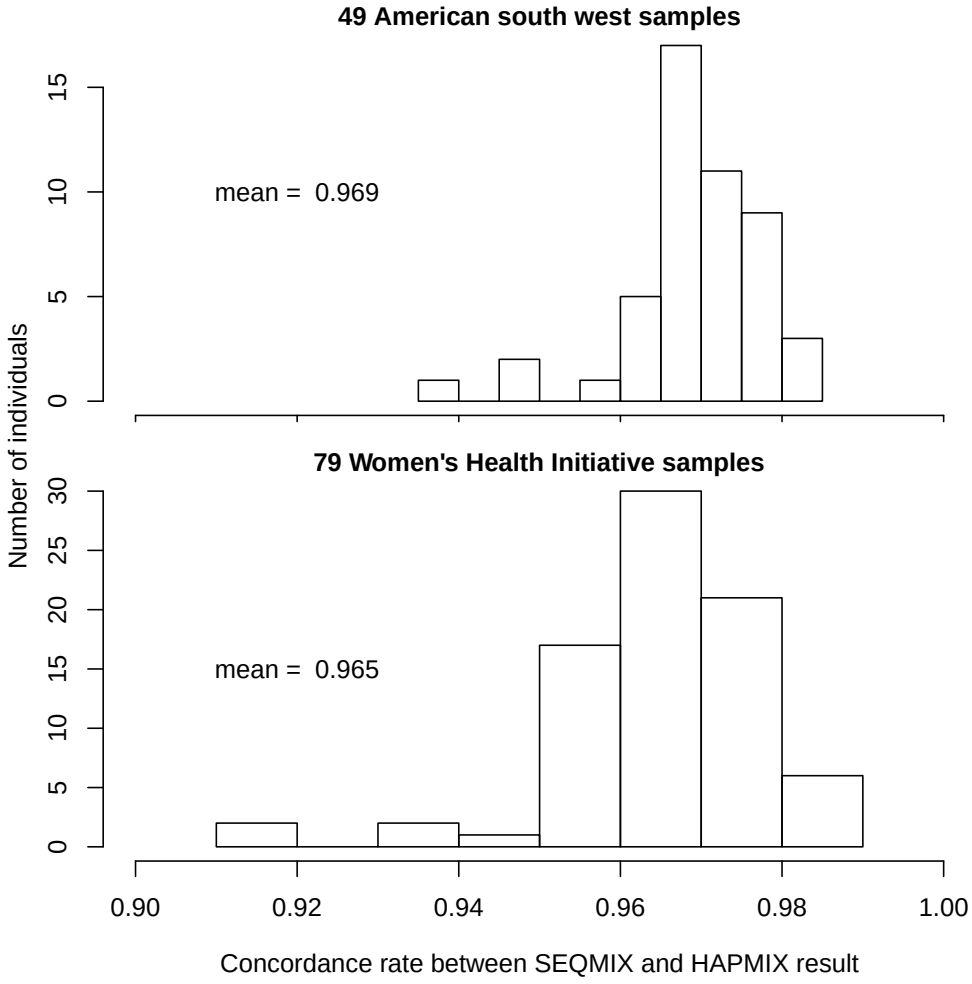

Figure 3.3: Distribution of concordance between SEQMIX and HAPMIX ancestry call results for two datasets: The top is from the 49 African American samples from the 1,000 Genome Project; the bottom is from the 79 African American samples from the Women's Health Initiative cohort.

Table 3.3: Pattern of coding variation in regions with diploid ancestry calls given by both SEQMIX and HAPMIX in the three ancestry categories: both haploids are of European ancestry; both are of African ancestry and one of each kind. Heterozygosity per kilobase, number of nonsynonymous sites per megabase and number of loss-of-function sites (annotated as stop or splice) per megabase are calculated for the 49 ASW samples.

	Exome size (Mb)		Heterozygosity (per kb)	
Ancestry	SEQMIX	HAPMIX	SEQMIX	HAPMIX
E/E	1.24	1.16	0.42	0.43
E/A	10.06	10.26	0.55	0.55
A/A	16.76	16.65	0.53	0.53
	# of nonsynonymous heterozygous sites (per Mb)		# of loss-of-function heterozygous sites (per Mb)	
Ancestry	SEQMIX	HAPMIX	SEQMIX	HAPMIX
E/E	190.78	189.61	1.88	1.74
E/A	246.81	248.23	2.13	2.15
A/A	239.24	237.33	2.01	2.22

sample has a mean on coverage 87 and an off target mean coverage 1.2. To run HAPMIX, we used the GWAS genotype data for these 79 samples as input, the $1,000$ Genomes Project European and African data as reference haplotypes and the interpolated genetic distance from IMPUTE as recombination rates.

These set of individuals have fractions of whole genome European ancestry varying between 2.1% to 53.5% based on result from SEQMIX. The corresponding numbers are 2.5% and 53.2% from HAPMIX results. Again, the two estimates are highly similar, with the squared Pearson correlation $> 99.7\%$.

Moreover, we present the whole genome concordance rate in the bottom panel of Figure 3.3. We see that the concordance rate varies between 0.91 and 0.99, with a mean 0.97. The squared correlation between the ancestries called by HAPMIX and SEQMIX has a median of 0.88, a mean of 0.85 and standard error 0.09 with the first quartile at 0.83 and third quartile at 0.91. These numbers indicate that the ancestry call from using off targeted reads using SEQMIX is again very similar to that from applying HAPMIX on the GWAS genotype arrays.

3.5 Discussion

We have described our method SEQMIX that uses off-target sequence reads from exome sequencing experiments to accurately infer whole genome ancestry in admixed samples. Simulations and real data analyses have shown that SEQMIX gives accurate local ancestry decompositions when the off target coverage is as low as 0.5x. SEQMIX should be useful for studies with exome or targeted sequence data of admixed populations, either for population history analysis or disease gene mapping.

Recent technology advancements have led to the emergence of sequence data and its successful application in understanding population genetics and disease gene mapping (*Bamshad et al.* 2011). Recent work has also shown that ultra low coverage sequencing can effectively replace genotyping technology for much reduced cost (*Pasaniuc et al.* 2012). Our method, at this crucial time, demonstrates that low coverage sequence data is effective at determining the local ancestries for admixed individuals.

While our method SEQMIX can provide very accurate local ancestry estimates with extremely low coverage sequence data, it also has some disadvantages. For examples, it requires allele frequencies and LD information from the reference populations. Future work would be to eliminate these requirements by incorporating the clustering algorithm in LAMP (*Sankararaman et al.* 2008) or relying on an initial PCA step (*Bryc et al.* 2010a). However, in applying these methods, we should take into consideration that there are many sites with no sequence reads. Moreover, in simulations, we have shown that ancestry blocks on the order of several megabases could be correctly estimated at ~ 0.5x off target mean coverage. Such blocks are typical for African American since the admixture event for this population happened very recently. For populations that were admixed in ancient history such as the Uyghur population in China, more work is needed to understand whether the typical current off target coverage (~ 0.5x) is sufficient. In developing SEQMIX, we focused our effort on decomposing ancestry for two-way admixture. In fact, our framework could be naturally extended to multi-way admixture. SEQMIX essentially models unlinked markers by incorporating a LD pruning data processing step. Future work can be done to extend our framework to model haplotypes to describe ancestry better. Another possible extension of SE-

QMIX includes integrating both GWAS array data and off target sequence data for more accurate ancestry modeling.

CHAPTER IV

The Benefits of Using Genetic Information to Design Prevention Trials

4.1 Abstract

Clinical trials for preventative interventions are complex and costly endeavors focused on identifying individuals likely to develop disease in a short timeframe, randomizing these individuals to different treatment groups, and then following them over time to capture disease onset events. In these prevention trials, statistical power is governed by the rate of disease events in each group and cost is dominated by randomization, treatment and follow-up. Strategies that increase the rate of disease events by enrolling individuals with high disease risk can significantly reduce study size, duration and cost.

Comprehensive study of common, complex diseases has resulted in a growing list of robustly associated genetic markers. Here, we evaluate the utility - in terms of trial size, duration and cost - of enriching prevention trial samples by combining clinical information with genetic risk scores to identify individuals at greater risk of disease. We also describe a framework for utilizing genetic risk scores in these

trials and evaluating the associated cost and time savings.

Using type 1 diabetes (T1D), type 2 diabetes (T2D), myocardial infarction (MI) and advanced age-related macular degeneration (AMD) as examples, we illustrate the potential and limitations of using genetic data for prevention trial design. These diseases differ in their genetic architecture: many markers are robustly associated with T2D and MI but all have relatively small effects; loci with large effect sizes have been identified for T1D and AMD. Our results illustrate settings where incorporating genetic information could reduce trial cost or duration up to 70%, as well as other settings were potential savings are modest. Results are strongly dependent on the genetic architecture of the disease, but we also show these benefits should increase as the list of robustly associated markers for each disease grows.

4.2 Author summary

Large-scale genetic association studies have identified many markers that are robustly associated with a variety of complex traits, such as diabetes and cardiovascular disease. Together, these markers may help identify individuals at high risk of disease and help design shorter and more cost-effective trials of new interventions to prevent disease. We quantify the potential benefits - in terms of prevention trial size, duration and cost - of using genetic risk factors to help identify individuals at greater risk of developing type 1 and type 2 diabetes, AMD and myocardial infarction.

4.3 Introduction

Designing a randomized clinical trial for disease prevention is a complex and costly endeavor (*Dickson and Gagnon* (2004)). A key step is to identify individuals likely to develop the disease during the study. The cost of a prevention trial strongly depends on the rate of disease onset among participants: low rates of disease onset require large sample sizes or long trial duration to achieve adequate statistical power. Most primary prevention trials thus apply 'enrichment' strategies to recruit individuals at high risk of disease onset (*Cummings et al.* (2007); *Florez et al.* (2006); *Ridker et al.* (2008)). Such trial design strategies also have ethical benefits because only at-risk subjects are exposed to potential side effects of a novel intervention. Enrichment designs can also be used in other types of clinical trials, including treatment clinical trials contexts (see *Simon* (2008) for examples).

Now genetic markers have been robustly associated with many complex diseases, it is timely to explore how genetic information, in conjunction with clinical information, can be used in the design of prevention trials (*Burke and Psaty* (2007)). This question can be decomposed into two more specific questions: first, how can we accurately predict disease risk from genetic data; and second, how can we use predicted genetic risks to design more efficient prevention trials? The question of predicting genetic risk of complex diseases has recently been explored in various contexts (*Pharoah et al.* (2008); *Yang et al.* (2010); *Sanna et al.* (2009)).

Original attempts to utilize genetic information in trial design proposed using a small number of risk genotypes as discrete inclusion criteria. Examples include using BRCA1 and BRCA2 genotypes in breast cancer prevention trials (*King et al.* (2001)), APOE/e4 genotypes in Alzheimer's Disease prevention trials (*Cummings*

et al. (2007)), and, more generally, key markers identified by genomewide association studies (GWAS) (*Schork and Topol* (2010)). Here, we explore an extension of this concept that can incorporate hundreds or thousands of robustly associated genetic markers, using a quantitative 'genetic risk score' aggregated across markers (*Lin et al.* (2009)). The cost and duration of a prevention trial will depend on the prediction accuracy of the risk score and the threshold used to select eligible subjects.

To evaluate the benefits and limitations of using genetic risk prediction models, we compare the cost and duration of prevention trials in various scenarios, including trials using only clinical information and trials also using genetic information to identify high-risk subjects. To illustrate the issues, we consider current risk prediction models for four diseases: type 1 diabetes (T1D), type 2 diabetes (T2D), myocardial infarction (MI) and age-related macular degeneration (AMD). Through simulation, we show that aggregate risk scores are expected to help reduce cost of clinical trials, sometimes modestly (T2D, MI) and sometimes substantially (T1D, AMD). Re-analyzing existing experimental data, we further evaluate our model in the context of T2D and AMD. Finally, we evaluate the utility of biobanks where a large number of genotyped individuals make enrichment based on genetic information particularly cost effective.

4.4 Materials and methods

4.4.1 Framework of genetic enrichment trial for disease prevention

We consider a standard design framework for prevention trials as 'conventional prevention trials'. Eligibility criteria are assessed in potential trial participants

after they provide informed consent. Typically, this involves selecting individuals likely to develop the disease based on clinical risk factors, such as glucose levels for T2D (*Florez et al.* (2006)) and low-density lipoprotein (LDL) or C-reactive protein (CRP) levels for MI (*Ridker et al.* (2008)). Additional risk variables such as age, gender, or smoking history may also be incorporated into the criteria (Table 4.1). Eligible participants are randomized to different treatment arms and followed for a trial period as illustrated in Figure 4.1A. The treatment effect will be evaluated by comparing the frequency of disease onset between arms. The inclusion criteria capitalize on prognostic factors that 'enrich' disease onset among the trial subjects. Studying these individuals increases the number of disease onset events and thus reduces the sample size and the trial cost.

In genetic enrichment trials, the inclusion criteria further incorporate genetic information in a quantitative manner (Figure 4.1B). In such trials, a larger number of potential participants are screened to obtain a small fraction of individuals at higher disease risk. Consequently, the targeted participants will be at higher risk than those in conventional trials, and they will also be more likely to develop the disease during the trial period.

Examining the tradeoff between resources used in the screening stage and the trial stage is essential to optimize the efficiency of the trial. If the eligible criteria are too stringent, the number of potential participants to recruit and screen will be orders of magnitude larger than that of conventional trials, and the associated costs of screening will become a substantial portion of the total trial cost. On the other hand, too liberal criteria will fail to enrich the disease onset among the trial participants, diminishing the benefits of genetic screening.

Another type of possible enrichment trial builds upon a pool of potential

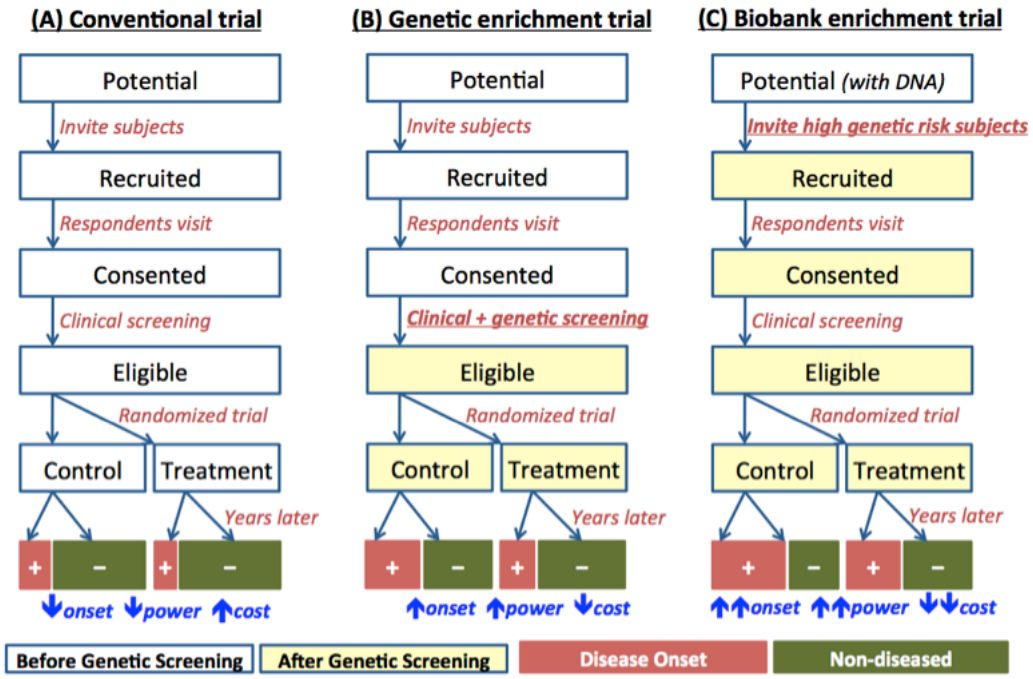

Figure 4.1: Frameworks of conventional and genetically enriched prevention trials. (A) Conventional prevention trial not utilizing genetic information, (B) Standard genetic enrichment trial only following up individuals at high genetic risk after genetic screening, and (C) Biobank-based enrichment trial where DNA information is available a priori and used for inviting individuals at the beginning of trial.

participants with genetic information readily available. Several large-scale DNA biobanks are currently being established with sample sizes up to hundreds of thousands patients (*Jayasinghe et al.* (2009)) with consent for genetic prescreening. Individuals found to be at high risk based on genetic risk factors determined from their banked genetic information would be prioritized for recruitment. Given a sufficiently large number of samples in the DNA biobank, this strategy makes it possible to identify an extremely small fraction of individuals at much higher risk than others without additional screening cost (Figure 4.1C).

4.4.2 Genetic risk model from known genetic risk variants

We consider a model of individual genetic risk based on markers known to be associated with disease traits with genome-wide significance. Typically an individual genetic risk score is calculated as a weighted sum of risk alleles (*Evans et al.* (2006))

$$\gamma(x) = \sum_i \gamma_i(x_i) = \sum_i x_i \log(\mathrm{OR}_i)$$

where x_i is 0, 1, or 2 copies of i-th risk alleles and OR_i denotes the odds ratio of i-th risk allele estimated from previous data.

The rate of disease onset in the prevention trial participants can be modeled as a logistic function of $\gamma(x)$, assuming that the effect size from risk variants remains the same in the trial population

$$Pr(d|x,z) = \frac{1}{1+\exp(-\mu-\gamma(x)-\delta z)} \tag{4.1}$$

where d denotes disease onset event during the trial period, μ is the intercept, and z is the binary indicator of randomization of treatment assignment with treatment

effect size of σ. μ and δ are iteratively adjusted so that $E[d|z=0]$ and $E[d|z=1]$ are equal to the rate of disease onset in the control and treatment arm, respectively. The receiver operating characteristics (ROC) of this genetic risk score can be obtained given the risk allele frequency and disease prevalence.

4.4.3 Simulation of genetic risk scores

We evaluated the performance of genetic enrichment trial designs using simulated genetic and phenotype data for four diseases. To simulate the genetic risk score using known genetic risk variants, we simulated genotypes of a million individuals based on risk allele frequencies reported from published results. For each simulated individual, a genetic risk score is evaluated using the published effect size of each risk variant and the simulated genotypes. The likelihoods of the individual having disease were evaluated using equation (4.1). To account for the variability in the estimates of odds ratios, we calculate individual risk scores by sampling odds ratios from reported confidence intervals 100 times and then repeated the simulation procedure described above using each set of sampled risk scores. The very optimistic and very pessimistic estimates of ROC in Figure 4.1 use the upper and lower bound of 95% confidence intervals of odd ratio for each of the risk variant, respectively.

4.4.4 Risk model from known AUC values

More generally, individual disease risk can be estimated from genetic and clinical information independently or collectively. In particular, we consider combined genetic and clinical risk from cohort studies of AMD and T2D (*Seddon et al.* (2009); *Talmud et al.* (2010)). In these studies, an alternative measurement - the

area under the ROC curve (AUC) (*Hanley and McNeil* (1982)) - are reported. In fact, for each set of AUC value and disease prevalence, there is a unique normal risk $\gamma(x)$ as in equation (4.1).

In the analysis of empirical data in AMD, the published AUC values were adjusted for inclusion criteria of baseline grade 3 or greater using the following equation

$$\mathrm{AUC}_{adj} = \frac{\mathrm{AUC}_{org} - f_2}{1 - f_2}$$

where AUC_{org} and AUC_{adj} denotes original and adjusted AUC values, respectively, and f_2 is the proportion of individual with baseline grade 2 or less. We have assumed no individual with grade 2 developed the disease during the trial period, which is a reasonable approximation given that only 8 out of 454 (2%) individuals with baseline grade 2 developed advanced AMD throughout the trial. For the analysis of T2D empirical data, the AUCs were calculated using the real data.

In addition to the analysis of empirical data, AUC-based methods were also used in two hypothetical simulation settings where 25% and 50% of the known heritability in liability scale (*Wray et al.* (2010)) can be explained by known genetic variants. This can prospectively project the degree of enrichment using genetic factors that will be discovered in the future.

4.4.5 Estimation of sample size, trial cost, and trial duration

Given a threshold t for the genetic risk score, the expected fraction of individuals with disease onset events during the trial can be modeled as

$$\begin{aligned} \pi_C(t) &= E[d = 1|\gamma(x) \geq t, z = 0] \\ \pi_T(t) &= E[d = 1|\gamma(x) \geq t, z = 1] \end{aligned}$$

where $\pi_C(t)$ and $\pi_T(t)$ are the rates of disease onset in the control and treatment arms, respectively. Given a false positive rate α $(= 0.05)$ and power $1 - \beta$ $(= 0.8)$, the required per-treatment group sample size follows *Lachine* (1981)

$$\begin{aligned} n(t) &= \frac{4\bar{\pi}(t)(1 - \bar{\pi}(t))(Z_{1-\beta} + Z_{\alpha/2})}{(\pi_C(t) - \pi_T(t))^2} \\ \bar{\pi}(t) &= \frac{\pi_C(t) + \pi_T(t)}{2} \end{aligned}$$

Given per-sample clinical screening cost C_s, follow-up cost C_f, and proportion of eligible participants f_e, the cost of a conventional prevention trial is determined as

$$(\frac{C_s}{f_e} + C_f)n(-\infty) \tag{4.2}$$

where $n(-\infty)$ represents sample size of conventional trial (see Figure 4.1A).

For a genetic enrichment trial (see Figure 4.1B) with additional genetic screening cost C_g, assuming that clinical and genetic screening is performed simultaneously with clinical screening, the overall cost becomes

$$(\frac{C_s + C_g}{f_e Pr(\gamma(X) \geq t)} + C_f)n(-\infty) \tag{4.3}$$

The reduction in years of trial at fixed sample size is iteratively estimated by modeling the disease progression rate as a function of trial duration, under the simplifying assumption that the rate of disease onset is constant over the course of trial period.

4.5 Result

We consider the three types of randomized two-arm primary prevention trials illustrated in Figure 4.1: (1) **conventional prevention trials** that screen potential participants using eligibility criteria based on a set of clinical variables, (2) **genetic enrichment prevention trials** that screen participants using clinical variables and genetic risk factors, and (3) **biobank enrichment prevention trials** that identify potential participants with high genetic risk scores prior to clinical screening. While we first evaluate the benefits of using genetic information using simulations, we later use empirical data from previous studies, such as the Age-Related Eye Disease (ARED) (*Seddon et al.* (2009)) and Whitehall II (*Talmud et al.* (2010)) studies, to account for potential overlap between clinical and genetic risk factors.

4.5.1 Effect of known risk variants on disease liability

We first evaluated the potential ability of GWAS variants to identify at risk individuals using simulations. We considered risk variants identified by large-scale meta-analyses for T1D (*Barrett et al.* (2009)), T2D (*Voight et al.* (2010)), MI (*Myocardial Infarction Genetics Consortium* (2009)) and AMD (*Chen et al.* (2010)) as robust genetic associations (see Table 4.1). Using published risk allele

frequencies and effect sizes, individual genetic risk scores were simulated assuming an additive model (See Methods for details). Figure 4.2 illustrates the distribution of the genetic risk score for individuals with and without disease for each trait. The distributions of the genetic risk scores in individuals with and without disease are very similar for T2D and MI but quite different for AMD and T1D, where a number of loci with large effect sizes have been described. These genetic risk profiles depend on current knowledge of the genetic architecture of each disease and can also be summarized as Receiver Operating Characteristics (ROC) curves that describe our ability to distinguish individuals with and without disease using genotypes. In addition to predictions based on published effect size estimates, ROC summarized in Figure 4.3 also include predictions that account for uncertainty in published effect sizes (Figure 4.3).

We next used simulations to predict the relative prevalence of disease in individuals with high and low genetic risk scores (Table 4.2). For AMD and T1D, we estimate that selecting individuals with genetic risks in the top decile would result in a $\sim 3-5$ fold increase in disease prevalence. Selecting individuals with genetic risks in the top percentile would result in a $\sim$ 5- to $\sim$ 13-fold increase in prevalence. For T2D and MI, $\sim 1.5-2$ fold increases in disease prevalence were expected among individuals with in the top decile, whereas $\sim 2-3$ fold increases in risk were expected among individuals with risks in the top percentile of genetic risk.

	Disease Trait	T2D	AMD	T1D	MI
Current Genetic Knowledge	Population Prevalence	3.0%	11.8+%	0.54%	4.0%
	Sibling recurrence risk	3.5	2.2	13.7	3.2
	Heritability in liability scale*	0.60	0.68	0.86	0.71
	# Known risk variants	29	7	41	12
	Range of odds ratios per allele	1.06~1.37	1.31~4.31	1.05~5.49	1.13~1.28
	Heritability explained by known genetic risk variants	3~9%	46~59%	18~29%	1~5%
Example Prevention Trial	Treatment effect	Thiazoli-dinedione	Zinc + Antioxidants	Oral Insulin	Statin
	Inclusion criteria**	IGT/IFG	≥3 baseline AMD grade	N/A**	N/A**
	Average annual rate of disease onset in control arm (Conventional trial)	8.7%	4.4%	2.1%	2.5%
	Average annual rate of disease onset in treatment arm (Conventional trial)	3.9%	3.2%	1.0%	1.5%
	Trial duration	3 years	5 years	4 years	5 years
	Clinical screening cost++	$1,500	$1,500	$1,500	$1,500
	Additional cost for genetic screening	$100	$100	$100	$100
	Trial cost per year per subject	$6,000	$3,500	$12,000	$6,000
	Additional per-subject cost for genetic screening	$100	$100	$100	$100

* Heritability is estimated from prevalence and sibling recurrence risk

**Inclusion criteria are applied only in the experimental data setting for T2D and AMD, not in the simulation-based studies

+ AMD prevalence from individuals with age 80 years or older

++ Screening and trial costs are assumed to take failure rate into account

Table 4.1: A summary of genetic and treatment information for four disease traits. The references we used for estimates of population prevalence are T2D: (*Das and Elbein* (2006)), AMD: (*Seddon et al.* (2005)), T1D: (*Hyttinen et al.* (2003)) and MI: (*Nora et al.* (1980)). The references we used for GWAS or the meta analysis results are listed as T2D: (*Voight et al.* (2010)), AMD (*Chen et al.* (2010)), T1D (*Barrett et al.* (2009)), and MI (*Myocardial Infarction Genetics Consortium* (2009))

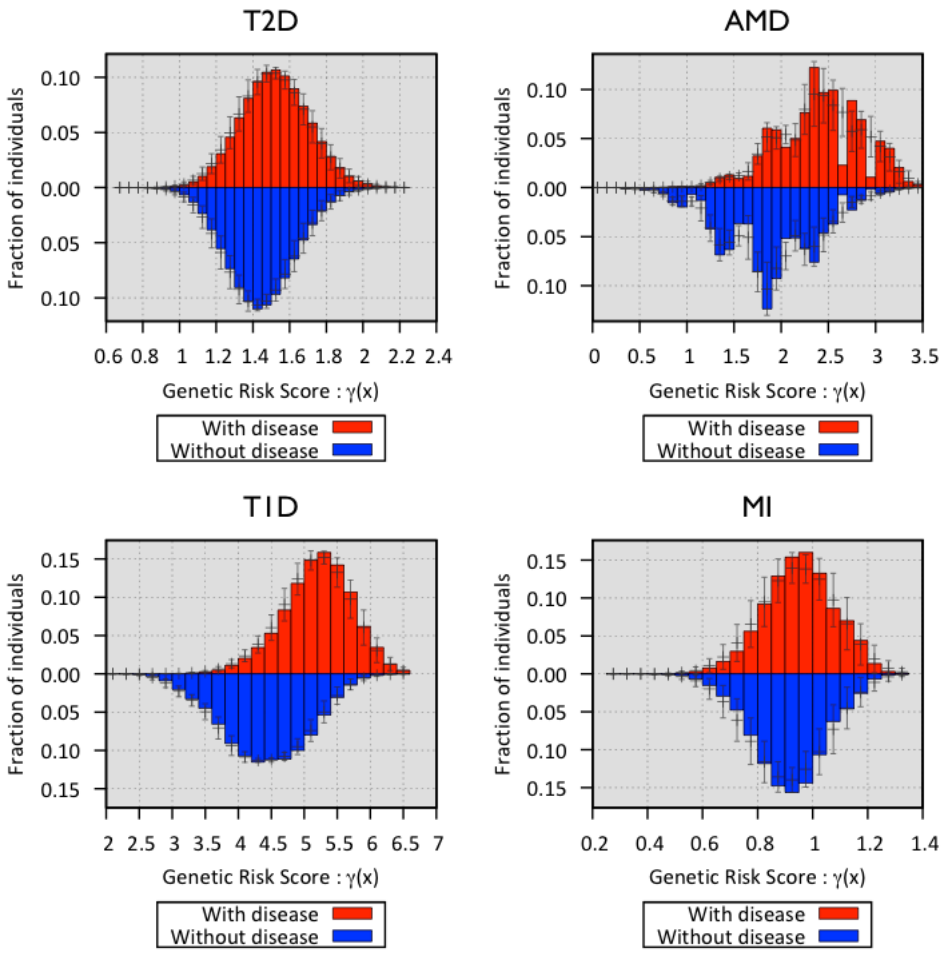

Figure 4.2: Distribution of genetic risk scores from currently known risk variants for four disease traits. The x-axis represent the genetic risk score with respect to the individuals with the lowest risk genotypes. The y-axis represents the fraction of individuals with disease based on their risk score. The 95% confidence intervals account for variations in the odds ratio estimates.

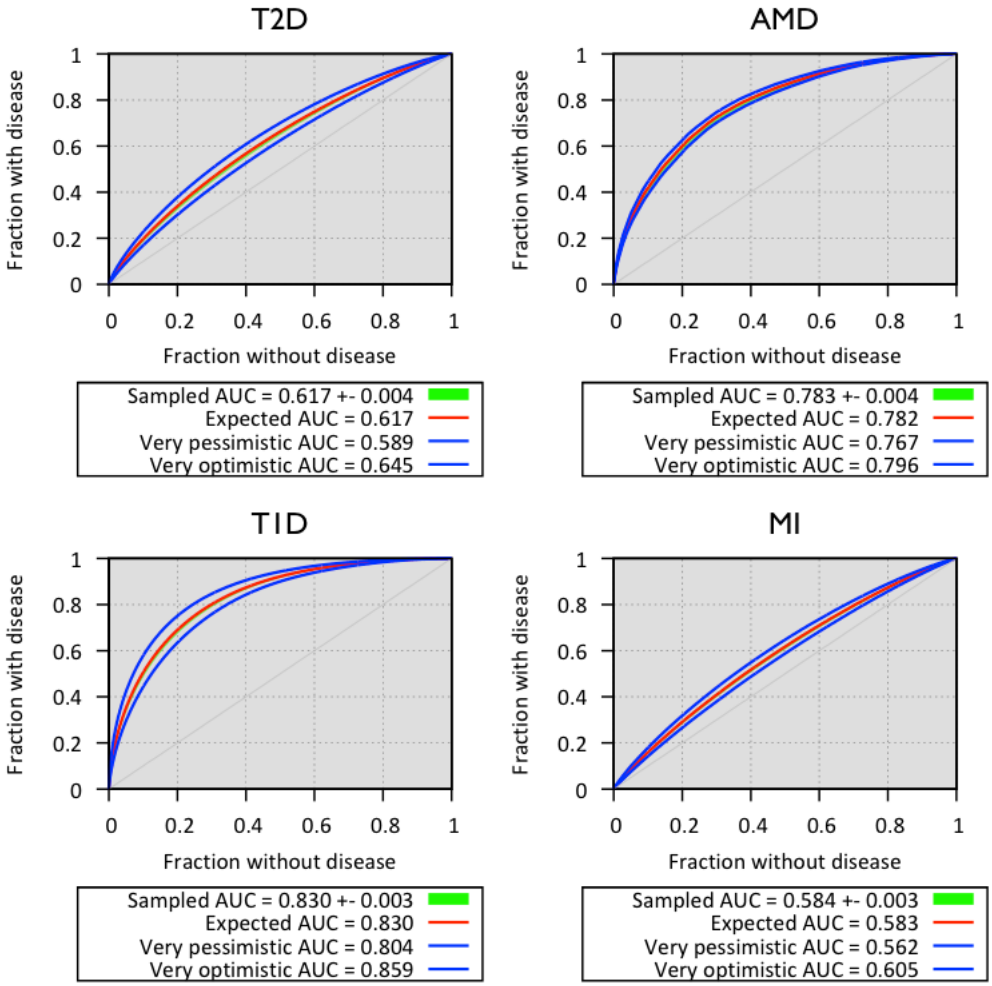

Figure 4.3: Receiver Operating Characteristics curve of the genetic risk score from known risk variants. Expected AUC represent area under the ROC curve using expected odds ratio. The sampled AUC is calculated from 100 sets of sampled odds ratios accounting for confidence intervals (CIs). Very pessimistic and optimistic AUC is computed from lower and upper bound of 95% CI of odds ratio from each SNP, respectively.

Genetic risk threshold	Fold enrichment from baseline disease prevalence (in parenthesis)*			
	T2D (3.0%)	AMD (11.8%)	T1D (0.54%)	MI (4.0%)
Top 50% genetic risk	1.32	1.65	1.84	1.23
Top 20% genetic risk	1.67	2.63	3.42	1.43
Top 10% genetic risk	1.92	3.36	4.96	1.57
Top 5% genetic risk	2.17	4.07	6.82	1.69
Top 2% genetic risk	2.50	4.63	9.78	1.83
Top 1% genetic risk	2.74	4.81	12.42	1.94

* The ratio of prevalence in the individuals with top genetic risk to the baseline prevalence in Table 1

Table 4.2: Disease liability explained by currently known risk variants.

4.5.2 Utility of known risk variants in efficient design of prevention trials

Next, we estimated the utility of genetic risk scores for trial design. We considered prevention trials for T1D, T2D, MI and AMD. In each case, we modeled treatment effect, trial cost, and durations according to previous studies (Table 4.1). These simulations made two important simplifying assumptions. First, since genetic association studies typically report the impact of individual risk alleles on prevalence (rather than incidence), we assumed that the impact of genetic variants on disease incidence rates and on prevalence would be the same. Second and more importantly, we also assumed that genetic risk scores and clinical covariates would be associated with disease risk independently. To the extent that clinical variables mediate the impact of genetic variants on disease risk, this assumption will lead to optimistic predictions of performance for trials that use both genetic and clinical covariates for enrichment. Our analysis of prospective data generated by the AREDS (on AMD) (*Seddon et al.* (2009)) and Whitehall (on T2D) (*Talmud et al.* (2010)) studies overcomes these limitations.

In our simulations, as individuals with higher risk were targeted, the incidence

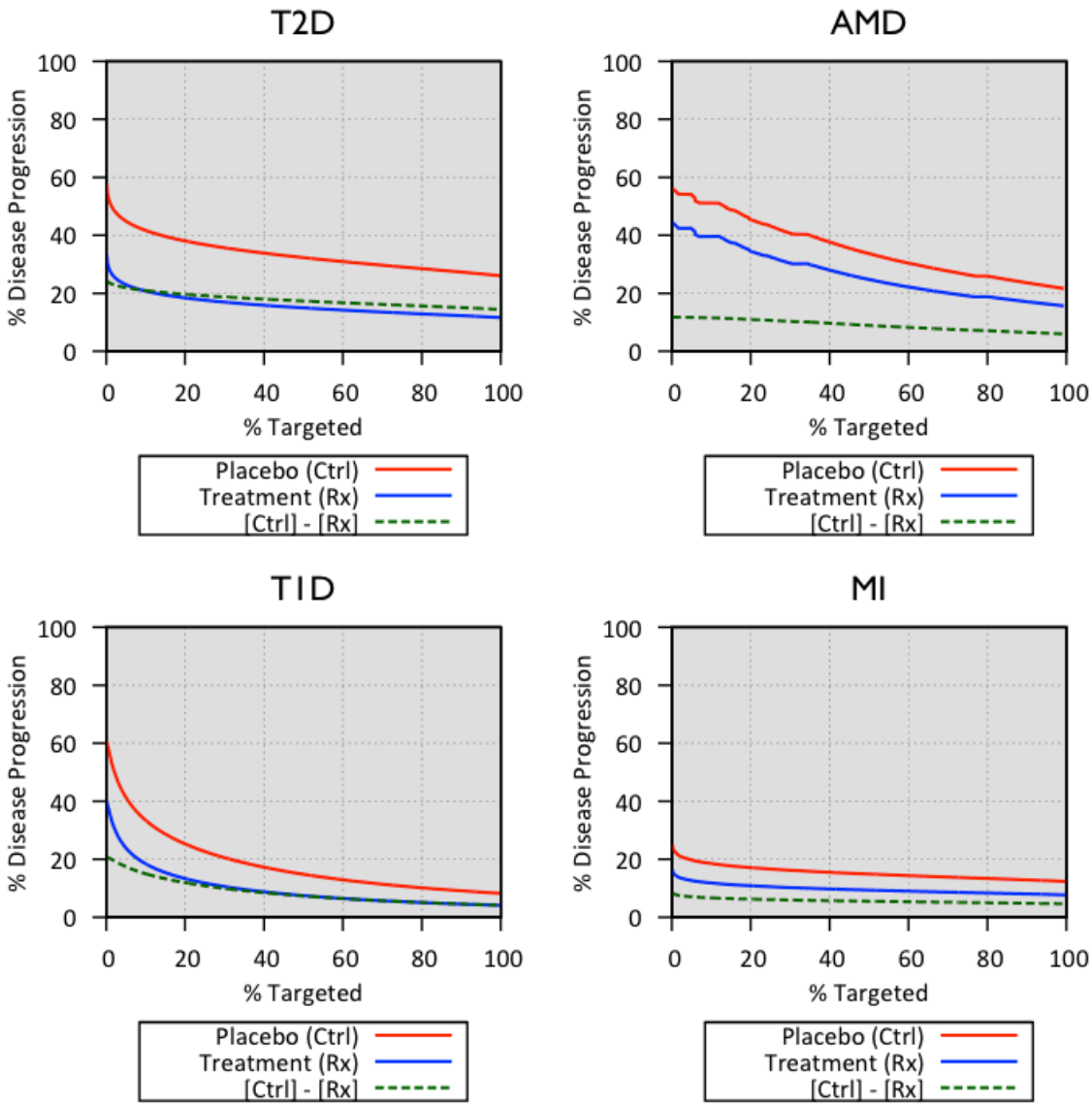

Figure 4.4: Changes in disease progression rate by the threshold of genetic risk score. The x-axis represents the targeted proportion of individuals at high genetic risk, and the y-axis represents the proportion of individuals with disease onset within 3 years of trial period in placebo (Ctrl) and Treatment (Rx) group. Dashed lines represent treatment effect as the differences between two progression rates.

of disease gradually increased in both treatment arms, whereas the treatment effect size only slightly increases (Figure 4.4). This increase in the disease incidence translates into reduced sample size requirements (Figure 4.5). At the same time, large increases in on-trial disease incidence require progressively larger samples to be screened for clinical and genetic risk factors, increasing screening costs. The optimal trial cost is determined by balancing these two tradeoffs. As shown in Table 4.3 and Figure 4.5, our simulation suggests cost savings up to 11% for T2D, 40% for AMD, 67% for T1D, and 13% for MI are possible when genetic enrichment is used to complement clinical risk factors. For a fixed sample size, genetic enrichment can reduce trial duration by 24% for T2D and MI, and $40-62\%$ for AMD and T1D.

4.5.3 Evaluation with experimental data

To complement these simulations, we applied our enrichment trial framework to longitudinal datasets documenting incidence of two specific diseases - AMD and T2D as well as clinical and genetic risk factors. This evaluation removes the simplifying assumptions required in our simulations.

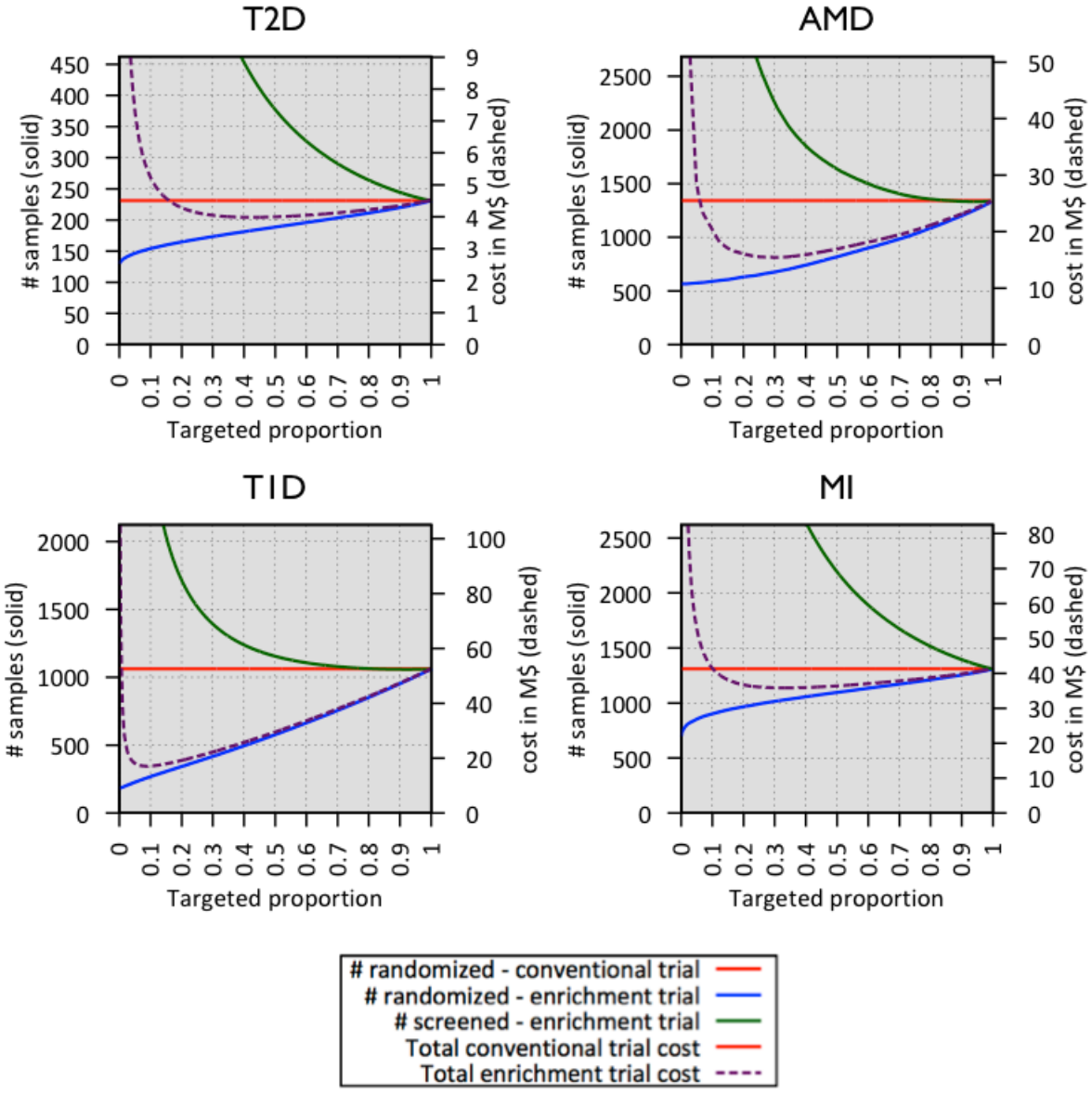

Figure 4.5: Sample size and total cost of genetically enriched prevention trials using currently known risk variants. X-axis represents the targeted proportion of individuals at high genetic risk, and the left y-axis, corresponding to solid lines, represents sample size for a conventional trial (red), on-trial sample size for a genetic enrichment trial (blue), and screening sample size for a genetic enrichment trial (green). The right y-axis, corresponding to dashed lines, represents the total cost of the genetic enrichment trial given targeted proportion.

Disease Trait	Trial Design	Optimized Trial Cost (Fixed trial duration)				Reduced trial duration (Fixed # subjects)			
		%Targeted Subjects	Trial # Subjects	Trial Duration	Total Cost ($)	%Targeted Subjects	Trial #Subjects	Trial Duration	Total Cost ($)
T2D	Conventional Trial	100%	231	3.0 yrs.	4.5M	100%	231	3.0 yrs.	4.5M
T2D	Genetic Enrichment Trial*	43%	184	3.0 yrs.	4.0M	20%	231	2.3 yrs.	5.1M
T2D	Biobank Enrichment Trial**	1%	136	3.0 yrs.	2.7M	1%	231	2.1 yrs.	3.3M
AMD	Conventional Trial	100%	1,342	5.0 yrs.	25.5M	100%	1,342	5.0 yrs.	25.5M
AMD	Genetic Enrichment Trial	31%	680	5.0 yrs.	15.4M	20%	1,342	3.0 yrs.	24.7M
AMD	Biobank Enrichment Trial	1%	565	5.0 yrs.	10.7M	1%	1,342	3.0 yrs.	16.1M
T1D	Conventional Trial	100%	1,061	4.0 yrs.	52.5M	100%	1,061	4.0 yrs.	52.5M
T1D	Genetic Enrichment Trial	9%	260	4.0 yrs.	17.1M	20%	1,061	1.5 yrs.	27.4M
T1D	Biobank Enrichment Trial	1%	190	4.0 yrs.	9.4M	1%	1,061	1.1 yrs.	16.1M
MI	Conventional Trial	100%	1,309	5.0 yrs.	41.2M	100%	1,309	5.0 yrs.	41.2M
MI	Genetic Enrichment Trial	34%	1,032	5.0 yrs.	35.8M	20%	1,309	3.8 yrs.	40.6M
MI	Biobank Enrichment Trial	1%	771	5.0 yrs.	24.3M	1%	1,309	3.2 yrs.	26.9M

* In genetic enrichment trials, the cost-optimizing fraction of targeted samples subjects is selected for determining reduced trial cost, and 20% of targeted samples subjects is selected for determining reduced trial duration.

** In biobank-based enrichment, 1% of targeted samples subjects is assumed to determine trial cost and sample size reduction.

Table 4.3: Sample size, cost, and trial duration of enrichment trials, simulated from published GWAS risk variants.

Because both clinical and genetic risk scores are available in this empirical setting, to precisely evaluate the additional benefits of genetic information, here we consider (1) conventional trials following up all eligible participants, (2) prevention trials focusing on individuals with high clinical risk scores based on clinical, demographic, and environmental variables, and (3) prevention trials focusing on individuals with high combined risk scores, incorporating both genetic and clinical risks.

Age-related macular degeneration

A published cohort study of 1,446 individuals at high risk of advanced AMD allows us to investigate our framework for this setting (*Seddon et al.* (2009)). Participants were assayed for known genetic risk variants in addition to demographic and environmental risk variables - age, gender, education, smoking history, and baseline AMD grade. In total, 19% (279) of the subjects developed advanced AMD (including unilateral and bilateral, and dry and wet types of advanced AMD), within 6.3 years of entering the study. The advantage of combining clinical and genetic risk compared to clinical risk only is reflected in area under the ROC curve (AUC) (*Rosner and Glynn* (2009)) statistics. A predictive model based on clinical variables alone resulted in AUC statistic of 0.757, while a predictive model using combined genetic and clinical variables resulted in an AUC statistic of 0.821. Among all risk variables considered, the baseline AMD grade was the strongest predictor of advanced AMD. Among the 454 individuals with a baseline low AMD grade of 2, only 8 (2%) of them developed advanced AMD during the trial period; in contrast, among 992 individuals with a baseline high AMD grade of $> 2,271$ (27%) developed advanced AMD during the trial period. To mimic a realistic scenario for an AMD prevention trial from the cohort study, we considered a pre-

vention trial using baseline grade ≥ 3 as inclusion criteria. We estimate that in this subset of individuals, the AUC would be 0.637 using only clinical predictors, and 0.743 with genetic and clinical predictors.

Based on these adjusted AUCs and the reported rate of disease onset for each of the treatment groups (*Seddon et al.* (2005)), we estimated the sample size requirements, trial cost and duration for evaluating the efficacy of zinc + antioxidant treatment (Table 4.4). Our results show that, compared to a conventional trial relying only on baseline AMD grade ≥ 3 as the inclusion criterion, enrichment based on clinical risk scores from the demographic and environmental risk variables could reduce trial cost by up to 15%, either by reducing sample size requirements by 24% or by reducing the trial duration by 24% at a fixed sample size. Enrichment using both clinical and genetic factors, can reduce trial cost by up to 33%, either by reducing sample size requirements by 44% or by reducing trial duration by 36% - corresponding to a substantial efficiency gains beyond enrichment using only clinical characteristics.

Disease Trait	Trial Design	Optimized Trial Cost++ (Fixed trial duration)				Reduced trial duration++ (Fixed # subjects)			
		%Targeted Subjects	Trial # Subjects	Trial Duration	Total Cost ($)	%Targeted Subjects	Trial # Subjects	Trial Duration	Total Cost ($)
T2D	Conventional Trial+	100%	231	3.0 yrs.	4.51M	100%	231	3.0 yrs.	4.51M
T2D	Clinical-only Enrichment Trial*	29%	124	3.0 yrs.	2.92M	20%	231	1.9 yrs.	3.95M
T2D	Combined (Clinical+Genetic) Enrichment Trial**	28%	120	3.0 yrs.	2.84M	20%	231	1.9 yrs.	3.93M
AMD	Conventional Trial	100%	1,342	5.0 yrs.	25.5M	100%	1,342	5.0 yrs.	25.5M
AMD	Clinical-only Enrichment Trial*	41%	1,018	5.0 yrs.	21.8M	20%	1,342	3.8 yrs.	28.5M
AMD	Combined (Clinical+Genetic) Enrichment Trial**	30%	753	5.0 yrs.	17.2M	20%	1,342	3.2 yrs.	25.7M

+ Conventional trial only uses basic inclusion criteria – IFG or IGT status for T2D and baseline AMD grade ≥ 3.

* In clinical-only enrichment trial, the clinical risk is calculated with a multitude of clinical factors

** In combined enrichment trial, we combine the clinical and genetic information and use this combined risk in the trial design

++ Cost-optimizing fraction of targeted samples subjects are selected for determining reduced trial cost, and 20% of targeted subjects are selected for determining reduced trial duration.

Table 4.4: Sample size, cost, and trial duration of enrichment trials based on experimental results (*Seddon et al.* (2009); *Talmud et al.* (2010)).

While simulation-based estimates using GWAS-based effect size estimates and assuming independence of clinical and genetic risk factors suggested a potential 40% savings in prevention trial cost, this empirical analysis suggests a savings of 33% in prevention trial cost when combining genetic and clinical variables as risk predictors. In this case, both estimates are similar suggesting that the assumptions above do not qualitatively affect the conclusions of our simulation based analysis.

Type 2 diabetes

To empirically evaluate the efficiency of T2D prevention trials we used data from the Whitehall II prospective cohort study. This longitudinal study recruited a cohort of civil servants, 25 to 55 years old in central London, from $1985 - 1988$ and followed them until $2003 - 2004$. The detailed design and data analysis were reported previously (*Talmud et al.* (2010)). Among $5,535$ participants, we selected $1,916$ pre-diabetic subjects with either impaired glucose tolerance (IGT) ($7.8-11.0$ mmol/dL) or impaired fasting glucose (IFG) level ($5.6-6.9$ mmol/dL) in the initial phase or clinical examination to resemble subjects typically recruited in a type 2 diabetes prevention trial (*Florez et al.* (2006)).

Using the Framingham offspring T2D risk scores calculated only from clinical variables (*Wilson et al.* (2007)), genetic risk scores calculated using 20 robustly associated variants (*Voight et al.* (2010)), and risk scores calculated using both clinical and genetic factors, we evaluated different strategies for trial enrichment. Consistent with the previous study (*Talmud et al.* (2010)), we find that the genetic risk scores alone do not effectively predict onset of T2D in this cohort (AUC: 0.52). The Framingham T2D risk scores from clinical variables (AUC: 0.75) or combined risk scores (AUC: 0.76) were much more informative in predicting progression of diabetes among at risk individuals.

In this case, we estimate that clinical risk-score based enrichment trials can reduce the trial cost by 35%, sample size by 46%, or the trial duration by 37% (compared to conventional trials using only IFG/IGT status as eligibility criteria). We also estimate that, in this case, using combined risk scores that also include genetic information would result in negligible additional benefits (Table 4.4). This finding reflects our limited knowledge of the genetic variants contributing to T2D risk (mirrored in their low AUC contribution) and is more pessimistic that the estimate of an 11% cost saving from our simulations.

4.5.4 Biobank-driven prevention trial designs

We also simulated biobank-driven enrichment trials, which rely on a very large set of individuals for whom genetic information is stored in a DNA biobank and who have consented to being invited to participate in clinical trails. Given planned biobanking efforts targeting $> 100,000$ individuals, this approach may allow identification of individuals with very rare and very high-risk genotypes for modest screening cost. We estimated the sample size, trial cost, and possible reduction in trial duration when biobanks that were 100x larger than the planned on trial sample sizes. In this case, individuals in top percentile of genetic risk might be targeted (Table 4.3) and, except for very simple predictors like age, traditional clinical risk factors would be ignored.

We estimate that such biobank-driven enrichment strategies might reduce the trial cost by 41% for T2D, 58% for AMD, 82% for T1D and 41% for MI, when combined with screening for clinical risk factors. These estimates correspond to a range of 20% to 37% in cost savings beyond those available in standard genetic enrichment trials.

4.5.5 Prospect of improved genetic risk prediction

To assess the impact of future improvements in genetic risk predictions, we simulated enrichment prevention trials using hypothetical sets of risk variants that might explaining 25% or 50% of the heritability (*Wray et al.* (2010)) for the four disease traits (Table 4.5). In these simulations, the genetic enrichment prevention trials of T2D and MI are estimated to achieve cost and trial size savings similar to those available for AMD and T1D. These results suggest that more complete catalogs of disease risk alleles may substantially increase the potential utility of genetic information for trial enrichment.

4.6 Discussion

With rapid advances in high-throughput biological screening strategies, there is great hope that genetic information will enable the design of more efficient clinical trials and that further gains in efficiency may be provided by other genomic predictors of disease (such as transcript levels, epigenomic modifications and proteomic profiles). Here, we evaluated the potential benefits of using genetic information for designing prevention trials and derive a framework for estimating the potential cost savings when genetic information is used to identify at risk individuals for inclusion in a trial.

Our results demonstrate that focusing on individuals with high genetic risk may allow for reduced trial cost and duration. Currently, large benefits from genetic enrichment trials are likely to be limited to diseases, such as AMD or T1D, where variants accounting for a large fraction of the heritability have been identified. However, future advances in genetic knowledge (driven by sequencing

studies and other studies of rare variation, for example) should extend the utility of genetic enrichment trials to broader sets of complex diseases, including conditions for which genetic enrichment is currently unlikely to succeed, such as T2D or MI.

It is important to note that the value of genetic information is dependent on the clinical variables available and the populations and timescale of interest. Recent studies on the AMD risk assessment from AREDs subjects reported that the improvement in AUC due to addition of genetic factors is considerably lower when additional clinical variables such as the presence of large drusen, advanced AMD in one eye, and family history are considered (*Seddon et al.* (2011)). When these additional covariates were included, they report that overall AUC was considerably increased, from 0.73 to 0.88, and exclusion of genetic factors only marginally reduces the AUC from 0.88 to 0.87. Most importantly, adjusting for the clinical variables, the estimated hazard ratio for CFH and ARMS2 alleles was substantially reduced from 1.97 to 1.28 and from 2.21 to 1.56, respectively. This suggests that much of the genetic risk may be manifesting in the presence of strong clinical predictors, and when these clinical predictors are included in models of short-term risk, there is limited additional predictive value in including genetic risk factors.

Our observation that very large gains in efficiency are possible when DNA biobanks with genetic information on 100,000s of potential trial participants is available is particularly interesting. In this setting, trials can focus on individuals who carry very rare combinations of many risk alleles. For example, by focusing on individuals in the top 1% of the genetic risk of T1D, T2D, AMD or MI, we predict cost savings of 82%, 40%, 58% and 41%. If basic clinical information is also stored in the biobank, the potential efficiency gains will be even larger.

Disease Trait	Trial Design	Optimized Trial Cost (Fixed trial duration)				Reduced trial duration (Fixed # subjectamples)			
		%Targeted Subjects*	Trial # Subjects	Trial Duration	Total Cost ($)	%Targeted Subjects*	Trial #Subjects	Trial Duration	Total Cost ($)
T2D	Genetic Enrichment – 25% heritability	30%	131	3.0 yrs.	3.05M	20%	231	1.9 yrs.	4.51M
T2D	Genetic Enrichment - 50% heritability	25%	98	3.0 yrs.	2.40M	20%	231	1.6 yrs.	4.09M
AMD	Genetic Enrichment - 25% heritability	33%	828	5.0 yrs.	18.5M	20%	1,342	3.1 yrs.	26.6M
AMD	Genetic Enrichment - 50% heritability	27%	629	5.0 yrs.	11.9M	20%	1,342	2.5 yrs.	24.2M
T1D	Genetic Enrichment - 25% heritability	8%	242	4.0 yrs.	16.5M	20%	1,061	1.5 yrs.	27.3M
T1D	Genetic Enrichment - 50% heritability	7%	771	4.0 yrs.	10.7M	20%	1.061	1.1 yrs.	22.7M
MI	Genetic Enrichment - 25% heritability	16%	560	5.0 yrs.	22.4M	20%	1,309	2.6 yrs.	30.9M
MI	Genetic Enrichment - 50% heritability	14%	405	5.0 yrs.	16.8M	20%	1,309	2.1 yrs.	27.2M

* Cost-optimizing fraction of targeted subjects is selected for determining reduced trial cost, and 20% of targeted subjects is selected for determining reduced trial duration.

Table 4.5: Sample size, cost, and trial duration of enrichment trials, simulated from hypothetical risk variants explaining 25% and 50% of heritability.

Our cost models assume a fixed cost of screening and treatment. They do not allow for cost savings that may be possible in very large screening efforts; or, conversely, for cost increases that might result from the necessity of extending screening to additional sites. They also assume that genetic risk factors do not impact treatment efficiency although that may not always be the case. Interestingly, we note that the ratio of screening, genotyping and on-trial costs has a noticeable impact on the potential benefits of genetic information for trial design. Since genetic information potentially allows for smaller numbers of on-trial individuals, its benefits are particularly important when the on-trial costs are large. For our hypothetical AMD enrichment trials, an increase in on trial cost per subject from $\$3,500$ to $\$20,000$ would mean that an enrichment strategy combining genetic and clinical variables could enable a savings of 42% in cost (versus 33%).

Our simulations required important assumptions - particularly, the assumption that clinical and genetic risk factors are independent. For T2D and AMD, we were able to overcome this limitation by extending our analysis to also consider empirical samples that included information on disease incidence as well as clinical and genetic risk factors. While similar empirical assessments remain to be done for MI and T1D, we predict that the outcome for MI will be similar to that for T2D (where we conclude currently available genetic markers will typically have limited utility), we expect the situation for T1D might be more similar to that for AMD (where currently available genetic markers can already enable large cost savings). Future improvements in modeling will benefit from estimates of the performance of combined genetic and clinical risk scores in prospective studies.

Here, we focused on evaluating the utility of genetic information for enriching prevention trials. However, we expect that the combination of genetic information

and clinical trials will be a fertile area of research - including not just advances in trial design, but also opportunities to use genetic variants to understand the biology of drug response and adverse events. In cases where screening for clinical risk factors is laborious and expensive, genetic risk scores may be used as filter that focuses the clinical screening on at-risk individuals (an example might be antibody test response screening used to identify individuals at risk of developing T1D (*Orban et al.* (2009)). Finally, for common diseases where the genetic architecture is poorly understood, a proxy for a high genetic risk score might be the presence of an affected first degree relative, such as a parent or sibling.

Our model allows estimation of trial cost and duration in a variety of enrichment scenarios, including eligibility-criteria based on clinical factors, genetic factors or their combination. While we haven't investigated multi-arm trial designs, our work can model the utility of biobank-driven enrichment (where genetic information may be available for 100,000s of individuals) or of advances in genetic information. We will make the code enabling others to evaluate cost, sample size and time requirements for different trial designs available from our website.

4.7 Acknowledgements

We would like to thank Soumitra Ghosh for his insight into type 1 diabetes trials and Paul Newcombe and Linda McCarthy for their comments on the age-related macular degeneration models. This work was supported by the British Heart Foundation (grant numbers PG/07/133/24260, RG/08/008, SP/07/007/23671), Senior Fellowship to A.D.H. (grant number FS/2005/125).National Heart Lung and Blood Institute (grant number HL36310) for M.Kivimaki's and M.Kumari's contribu-

tions to this work; the Medical Research Council (grant number G0802432) by the Health and Safety Executive; the Department of Health; the National Institute on Aging in the United States (grant number AG13196); the Agency for Health Care Policy Research (grant number HS06516); the John D. and Catherine T. MacArthur Foundation Research Networks on Successful Midlife Development and Socio-economic Status and Health.

CHAPTER V

Conclusions

5.1 Summary

Clinical trials and genetics are two seemly disconnected fields in biomedical research. In this dissertation, we have described and implemented methods on efficient clinical trial design and accurate ancestry inference for exome sequence samples.

In Chapter 2, we derived the sample size formula for quadratic inference functions (QIF) in longitudinal design with dichotomous outcomes. We first introduced the method – QIF – which is an improvement of generalized estimating equations (GEE). We then presented the longitudinal logistic model and the detailed steps of sample size and power calculation for QIF and GEE for either known and unknown true correlation. We illustrated that QIF-based design demands less sample size than the GEE-based design. We further studied the property of both sample size formulas in relation to the number of follow-up visits and found that QIF is more robust than GEE. We recommend QIF as the choice of statistical method for longitudinal study designs.

In Chapter 3, we proposed our method SEQMIX that can use the off-target low coverage reads to accurately infer the local ancestry of exome sequenced admixed individuals. We simulated African American samples using European and African genomic templates from the 1,000 Genome project. By varying the off target mean coverage in the simulated African American sequence data, we verify that current exome sequence data with ∼1x off target mean coverage is sufficient in deciphering the ancestry blocks for populations that were admixed in recent history. Using two sets of real African Americans data, we further validate that using SEQMIX on exome data can derive ancestry estimates that are very similar to those from using high density genotype array data.

In Chapter 4, we connect the two research areas clinical trial and genetics studies by quantifying the benefits of using genetic information to design prevention trial. We evaluate the utility – in terms of trial size, duration and cost – of enriching prevention trial samples by combining clinical information with genetic risk score to identify individuals at greater risk of disease. We use type I diabetes (T1D), type 2 diabetes (T2D), myocardial infarction (MI) and age related macular degeneration (AMD) as examples to illustrate the potential and limitations of using genetic data for prevention trial designs. These are diseases with different genetic architecture: many markers are robustly associated the T2D and MI but all have relatively small effect sizes; markers with large effect sizes have been identified for T1D and AMD. Our results illustrate settings where incorporating genetic information could reduce trial cost or duration considerably as well as scenarios where potential savings are negligible. This work shows that the benefit of genetic information on clinical trials is highly dependent on the genetic architecture. We also project that the benefits should increase as the list of markers grows.

5.2 Relevance and Future work

This dissertation covers statistical methods on two distinct areas: efficient longitudinal clinical trial design and whole genome sequencing studies. Our sample size calculation based on QIF produces a more efficient longitudinal trial than one based on GEE. We have provided the software QIFSAMS for free download at `http://www-personal.umich.edu/~pxsong`. This calculator will be helpful for trial practitioners in seeking reduction of cost and resources for the trials. Our method SEQMIX, which takes advantage of the extremely low coverage off target reads for accurate local ancestry inference, will be useful for anyone undertaking exome or targeted sequencing studies, either to understand population history, or to conduct disease gene mapping. Our trial enrichment framework provides a general framework that can be used to evaluate the benefit of using genetic scores in the clinical trial designs. This model can also be effectively used to assess other types of scores either from combining different biomarkers or being derived from ROC curves.

Future extension includes applying QIF for genetic studies of families as it can be potentially useful in analyzing correlated phenotypes collected in genetic studies. SEQMIX can be extended to incorporate high density genotype data for better resolution of the ancestry estimates. SEQMIX currently requires allele frequencies and LD estimates for the reference populations and we could further explore the options of eliminating this external information. In terms of incorporating genetic score for clinical trial design, we did not take the correlation between genetic and clinical variables into account. This shall be implemented in the future by modifying the logistic model to incorporate such correlations.

BIBLIOGRAPHY

BIBLIOGRAPHY

Bamshad, M., S. Ng, A. Bigham, H. Tabor, M. Emond, D. Nickerson, and J. Shendure (2011), Exome sequencing as a tool for mendelian disease gene discovery, *Nature Reviews Genetics*.

Barrett, J., D. Clayton, P. Concannon, and et al (2009), Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes, *Nature Genetics*, *41*, 703 – 707.

Bentley, D., et al. (2008), Accurate whole human genome sequencing using reversible terminator chemistry, *Nature*, *456*.

Bilguvar, K., A. Kemal, A. Louvi, K. Kwan, and M. Choi (2011), Whole-exome sequencing identifies recessive wdr62 mutation in severe brain malformations, *Nature*, *467*, 207 – 210.

Bryc, K., A. Auton, M. Nelsen, J. Oksenberge, and S. Hauser (2010a), Genome-wide patterns of popoulation structure and admixture in west africans and african americans, *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 786 – 791.

Bryc, K., C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C. D. Bustamante, and H. Oster (2010b), Genome-wide patterns of population structure and admixture among hispanic/latino populations, *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 8954 – 8961.

Burke, W., and B. Psaty (2007), Personalized medicine in the era of genomics, *Journal of American Medical Association*, *298*, 1682 – 1684.

Chen, W., D. Stambolian, A. Edwards, and et al (2010), Genetic variants near timp3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration, *Proc Natl Acad Sci U S A*, *107*, 7401 – 7406.

Choi, M., U. Scholl, W. Ji, T. L. ands Irina Tikhonova, P. Zumbo, A. Nayir, and A. Bakkaloglu (2009), Genetic diagnosis by whole exome capture and massively

parallel dna sequencing, *Proceedings of the National Academy of Sciences of the United States of America*, *106*(45), 19,096 – 10,101.

Cirulli, E., and D. Goldstein (2010), Uncovering the roles of rare variants in common disease through whole-exome sequencing, *Nature Review Genetics*, *11*, 415 – 425.

Cummings, J., R. Doody, and C. Clark (2007), Disease-modifying therapies for alzheimer disease: challenges to early intervention, *Neurology*, *69*, 1622 – 1634.

Das, S., and S. Elbein (2006), The genetic basis of type 2 diabetes, *Cellscience*, *2*, 100 – 131.

Demidenko, E. (2006), Sample size determination for logistic regression revisited, *Statistics in Medicine*, *26*(18).

Dickson, M., and J. Gagnon (2004), Key factors in the rising cost of new drug discovery and development, *Nature Review Drug Discovery*, *3*, 417 – 429.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002), *Analysis of Longitudinal Data*, Oxford University Press.

Evans, D., J. Marchini, A. Morris, and L. Cardon (2006), Two-stage two-locus models in genome-wide association, *PLoS Genetics*, *2*.

Florez, J., K. Jablonski, N. Bayley, T. Pollin, and P. Debakker (2006), Tcf7l2 polymorphisms and progression to diabetes in the diabetes prevention program, *New England Journal of Medicine*, *355*, 241 – 250.

Freedman, M., C. Haiman, N. Patterson, G. McDonald, and A. Tandon (2006), Admixture mapping identify 8q24 as a prostate cancer risk locus in african-american men, *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 14,068 – 14,073.

Gravel, S. (2012), Population genetics model of local ancestry, *Genetics*.

Hanley, J., and B. McNeil (1982), The meaning and use of the area under a receiver operating characteristic (roc) curve, *Radiology*, *143*(29).

Hansen, L. (1982), Large sample properties of generalized method of moments estimators, *Econometrica*, *50*, 1029 – 54.

Hsieh, F., D. Bloch, and M. Larsen (1998), A simple method of sample size calculation for linear and logistic regression, *Statistics in Medicine*, *17*(14), 1623 – 34.

Hyttinen, V., J. Kaprio, L. Kinnunen, and et al (2003), Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs: a nationwide follow-up study, *Diabetes*, *52*, 1052 – 1055.

Jayasinghe, S., A. Mishra, D. Van, and E. Kwan (2009), Genetics and cardiovascular disease: Design and development of a dna biobank, *Exp Clin Cardiol*, *14*, 33 – 37.

Jung, S.-H., and C. Ahn (2003), Sample size estimation for gee method for comparing slopes in repeated measurement data, *Statistics in Medicine*, *22*, 1305 – 1315.

King, M., S. Wieand, K. Hale, M. Lee, and et al (2001), Tamoxifen and breast cancer incidence among women with inherited mutations in brca1 and brca2: National surgical adjuvant breast and bowel project (nsabp-p1) breast cancer prevention trial, *Journal of American Medical Association*, *286*, 2251 – 2256.

Lachine, J. (1981), Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials*, *2*, 93 – 113.

Lambert, D. (2010), Zero-inflated poisson regression, with an application to defects in manufacturing, *Technometrics*, *34*, 1 – 14.

Li, H. (2011), A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*, *27*, 2987 – 2993.

Liang, K., and S. Zeger (1986), Longitudinal data analysis using generalized linear models, *Biometrics*, *73*, 13 – 22.

Lin, X., K. Song, and N. L. et al (2009), Risk prediction of prevalent diabetes in a swiss population using a weighted genetic score–the colaus study, *Diabetologia*, *52*, 600 – 608.

Lupski, J., J. Reid, C. Gonzaga-Jauregui, D. Deiros, and D. Chen (2011), Whole-genome sequencing in a patient with charcot-marie-tooth neuropathy, *The New England Journal of Medicine*, *362*, 1181 – 1191.

Maher, B. (2008), Personal genomes: the case of the missing heritability, *Nature*, *456*, 18 – 21.

Manolio, T., F. Collins, N. Cox, D. Goldstein, and L. Hindorfd (2009), Finding the missing heritability of complex diseases, *Nature*, *461*, 747 – 753.

Marchini, J., B. H. ands S Myers, G. McVean, and P. Donnelly (2007), A new multipoint method for genome-wide association studies by imputation of genotypes, *Nature Genetics*, *39*, 906 – 913.

Myocardial Infarction Genetics Consortium (2009), Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants, *Nature Genetics*, *41*, 334 – 341.

Ng, S., E. Turner, P. Roberson, S. Flygare, A. Bigham, and C. Lee (2009), Targeted capture and massively parallel sequencing of 12 human exomes, *Nature*, *461*(10).

Ng, S., A. Bigham, K. Buckingham, and M. Hannibal (2010a), Exome sequencing identifies mll2 mutation as a cause of kabuki syndrome, *Nature Genetics*.

Ng, S., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, and et. al. (2010b), Exome sequencing identifies the cause of a mendelian disorder, *Nature Genetics*, *42*(30 – 35).

Nora, J., R. Lortscher, R. Spangler, A. Nora, and W. Kimberling (1980), Genetic–epidemiologic study of early-onset ischemic heart disease, *Circulation*, *61*.

Orban, T., J. Sosenko, D. Cuthbertson, J. Krischer, and J. Skyler (2009), Pancreatic islet autoantibodies as predictors of type 1 diabetes in the diabetes prevention trial-type 1, *Diabetes Care*, *32*(2269 – 2274).

Overall, J. E., and S. Tonidandel (2004), obustness of generalize estimation equation (gee) test of significance against misspecification of the error structure model. john e. overall and scott tonidandel, *Biometrics*.

Pan, W. (2001a), Sample size and power calculation with correlated binary data, *Controlled Clinical Trials*, *22*, 211 – 227.

Pan, W. (2001b), Akaike's information criterion in generalized estimating equations, *Biometrics*, *57*, 120–125.

Pasaniuc, B., N. Rohland, P. McLaren, K. Garimella, and N. Zaitlen (2012), Extremely low-coverage sequencing and imputation increases power for genome-wide association studies, *Nature Genetics*, *44*, 631 – 635.

Patterson (2004), Methods for high-density admixture mapping of disease genes, *The American Journal of Human Genetics*, *74*(5), 979 – 1000.

Pharoah, P., A. Antoniou, D. Easton, and B. Ponder (2008), Polygenes, risk prediction, and targeted prevention of breast cancer, *New English Journal of Medicine*, *358*, 2796 – 2803.

Price, A., A. Tandon, N. Patterson, and et al (2009), Sensitive detection of chromosomal segments of distinc ancestry in admixed populations, *PLoS Genetics*.

Qu, A., B. G. Lindsay, and B. Li (2000), Improving generalised estimating equations using quadratic inference function, *Biometrika*, *87*(4), 823–836.

Rabiner, L. (1986), An introduction to hidden markov models, *ASSP Magazin, IEEE*, *3*, 4 – 16.

Reich, D., N. Patterson, P. DeJager, G. McDonald, and A. Waliszewska (2005), A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility, *Nature Genetics*, *37*, 1113 – 1118.

Ridker, P., E. Danielson, F. Fonseca, J. Genest, and A. Gotta (2008), Rosuvastatin to prevent vascular events in men and women with elevated c-reactive protein, *New English Journal of Medicine*, *359*, 2195 – 2207.

Rochon, J. (1998), Application of gee procedures for sample size calculations in repeated measures experiments, *Statistics in Medicine*, *98*(E2), 3247–3259.

Rosner, B., and R. Glynn (2009), Power and sample size estimation for the wilcoxon rank sum test with application to comparisons of c statistics from alternative prediction models., *Biometics*, *65*.

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin (2008), Estimating local ancestry in admixed population, *The American Journal of Human Genetics*, *82*, 290 – 303.

Sanna, S., B. Li, A. Mulas, C. Sidore, and H. Kang (2009), Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability, *PLoS Genetics*, *7*.

Schork, N., and E. Topol (2010), Genotype-based risk and pharmacogenetic sampling in clinical trials, *Journal of Biopharmaceutical Statistics*, *20*, 315 – 333.

Seddon, J., J. Cote, W. Page, S. Aggen, and M. Neale (2005), The us twin study of age-related macular degeneration: relative roles of genetic and environmental influences, *Arch Ophthalmol*, *123*.

Seddon, J., R. Reynolds, J. Maller, and et al (2009), Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables, *Invest Ophthalmol Vis Sci*, *50*, 2044 – 2053.

Seddon, J., R. Reynolds, Y. Yu, and M. Daly (2011), Risk models for progression to advanced age-related macular degeneration using demographic, environmental, genetic, and ocular factors, *Ophthalmology*, *118*, 2203 – 2211.

Seldin, M. F., B. Pasaniuc, and A. L. Price (2011), New approaches to disease mapping in admixed populations, *Nature Review Genetics*, *12*, 523 – 528.

Shriner, D., A. Adeyemo, E. Ramos, G. Chen, and C. N. Rotimi (2011), Mapping of disease-associated variants in admixed populations, *Genome Biology*, *12*(223).

Simon, R. (2008), Development and validation of biomarker classifier for treatment selection, *J Stat Plan Inference*, *138*(2), 308 – 320.

Smith, M., N. Patterson, J. Lautenberger, A. TrueLove, G. McDonald, and et al (2004), A high-density admixture map for disease gene discovery in african americans, *American Journal of Human Genetics*, *456*(74), 1001 – 1013.

Song, P. X. K. (2007), *Correlated Data Analysis*, NewYork, Springer.

Song, P. X.-K., Z. Jiang, E. Park, and A. Qu (2009), Quadratic inference functions in marginal models for longitudinal data, *Statistics in medicine*, *28*(29), 3683 – 3696.

Sundquist, A., E. Fratkin, C. Do, and S. Batzoglou (2008), Effect of genetic divergence in identifying ancestral origin using hapaa, *Genome Research*, *18*, 676 – 682.

Talmud, P., A. Hingorani, J. Cooper, and M. M. et al (2010), Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall ii prospective cohort study, *BMJ*, *340*.

Tang, H., M. Coram, P. Wang, X. Zhu, and N. Risch (2006), Reconstructing genetic ancestry blocks in admixed individuals, *The American Journal of Human Genetics*, *79*, 1 – 12.

Teer, J., and J. Mullikin (2010), Exome sequencing: the sweet spot before whole genomes, *Human Molecular Genetics*, *19*, 145 – 151.

Teerenstra, S., B. Lu, J. S. Preisser, T. van Achterberg, and G. F. Borm (2010), Sample size considerations for gee analyes of three-level cluster randomized trials, *Biometrics*.

The 1000 Genome Project Consortium (2010), A map of human genome variation from population-scale sequencing, *Nature*, *467*, 1067 – 1073.

The Women's Health Initiative Study Group (1998), Design of the women's health initiative clinical trial and observational study, *Controlled Clinical Trials*, *19*, 61 – 109.

Voight, B., L. Scott, V. Steinthorsdottir, and et al (2010), Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis, *Nature Genetics*, *42*, 579 – 589.

White, H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, *48*, 817–838.

Wilson, P., J. Meigs, L. Sullivan, C. Fox, D. Nathan, and R. D'Agostino (2007), Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study, *Arch Intern Med*, *167*.

Winkler, C. A., G. W. Nelson, and M. W. Smith (2010), Admixture mapping comes of age, *Annual Review Genomic Human Genetics*, *11*, 65 – 89.

Worthey, E., A. Mayer, G. Syverson, and D. Helbling (2011), Making a definitive diagnosis: Successful clinical application of whole exome sequecing in a child with intractable inflammatory bowel disease, *Genetics in Medicine*, *13*(3), 255 – 262.

Wray, N., J. Yang, M. Goddard, and P. Visscher (2010), The genetic interpretation of area under the roc curve in genomic profiling, *PLoS Genetics*, *6*(2).

Xu, S., W. Huang, J. Qian, and L. Jin (2008), Analysis of genomic admixture in uyghur and its implication in mapping strategy, *The American Journal of Human Genetics*, *82*, 883 – 894.

Yang, J., B. Benyamin, B. McEvoy, S. Gordon, and A. Henders (2010), Common snps explain a large proportion of the heritability for human height, *Nature Genetics*, *42*, 565 – 569.

Zhu, X., A. Luke, R. Cooper, T. Quertermous, and C. Hanis (2005), Admixture mapping for hypertension loci with genome-scan marker, *Nature Genetics*, *37*, 177–181.