Analysis of Familial Aggregation Using Recurrence Risk for Complex Survey Data

by Cong Wang

B.A. in Information Management and Information Systems, June 2011,
University of International Business and Economics
M.S. in Statistics, May 2013, The George Washington University

A Dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

January 19, 2018

Dissertation directed by

Zhaohai Li
Professor of Statistics

Barry I. Graubard
Senior Investigator, National Cancer Institute

The Columbian College of Arts and Sciences of The George Washington University certifies that Cong Wang has passed the Final Examination for the degree of Doctor of Philosophy as of October 24, 2017. This is the final and approved form of the dissertation.

Analysis of Familial Aggregation Using Recurrence Risk for Complex Survey Data

Cong Wang

Dissertation Research Committee:

Zhaohai Li, Professor of Statistics, Dissertation Co-Director

Barry I. Graubard, Senior Investigator, National Cancer Institute, Dissertation Co-Director

Qing Pan, Associate Professor of Statistics, Committee Member

Yan Ma, Associate Professor of Epidemiology and Biostatistics, Committee Member

Dedication

To my family.

Acknowledgments

First of all, I would like to thank my dissertation advisor and co-advisor, Professor Zhaohai Li and Dr. Barry I. Graubard for their continuous support, encouragement and guidance. I am deeply grateful for their valuable advice and insights, which go beyond the call of their advisor duties.

During my time at George Washington University, Prof. Li has set an example of excellence as a researcher, mentor and educator. He taught me to organize thoughts, research a problem and express ideas. During my time at NCI, Dr. Graubard has always been patient, supportive and helpful. His knowledge and enthusiasm for research lead us through numbers of difficulties. I profoundly appreciate Prof. Li's and Dr. Graubard's commitments to my growth and progress in my Ph.D. program.

I would also like to thank Professors Qing Pan, Yan Ma, Emre Barut and Dr. Katki Hormuzd for critical comments and valuable suggestions to my dissertation, and their positive support as my dissertation committee members. They have dedicated their time and effort to help me improve this document. Additionally, I am grateful to Professor Xiaoke Zhang for being my dissertation defense committee chair.

I also give thanks to all faculty and staff members of the Department of Statistics at George Washington University for the guidance and education I received throughout my master and doctoral programs.

My family deserves my final and most sincere expression of gratitude. I would like to thank my parents, Xiting Wang and Qi Shan for their continual support, selfless affection and dedication throughout not only my time in graduate school, but also through my entire life. My beloved husband, Chen Xing, has also provided love,

comfort and continuous encouragement.

I also want to thank other family members and family-like friends including Zhe Shi, Chen Chen, Yanjun Li, Aotian Yang, Yi Lu, Lulu Wang and those I fail to mention here.

Abstract of Dissertation

Analysis of Familial Aggregation Using Recurrence Risk for Complex Survey Data

Familial aggregation of a disease is important for studying possible genetic etiology of a disease. A popular and useful measure of family aggregation is recurrence risk. It is the probability that a family member of an affected individual is also affected. A problem when estimating recurrence risk is how to allow for varying the family sizes. The quadratic exponential model (QEM) has been proposed for estimating parameters for computing the recurrence risk when there is fixed family size in the sample. In this thesis, we use a marginal estimation approach for the QEM based on a pairwise composite likelihood to address varying family sizes. Household health surveys with (family) network sampling, which surveyed individuals report about disease status of themselves and specified relatives, has been shown to be useful for estimating prevalence of diseases and more recently for estimating recurrence risk of disease using nonparametric classical survey methods. Because these surveys have complex sample designs such as stratified multistage cluster designs with sample weighting for differential sample selection rates, this thesis extends the composite pairwise likelihood estimation and hypothesis of parameters of the QEM for simple random samples to data from these complex sample designs. In addition, the QEM is extended to simultaneously estimate and test parameters and recurrence risk for multiple family relationships, for comparing recurrence risk across family-level covariates (e.g., race) and utilizing propensity score weighting to adjust for confounding by individual-level covariates (e.g., age). Simulations are used to study the finite sample properties of the parameter estimation, variance estimation and level and power of hypotheses testing based on derived Wald and Quasi-score tests for these extended QEMs. Finally, our methods are illustrated using the 1976 National Health Interview Survey diabetes data set.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Chapter 1 introduces the overall background related to the familial aggregation and complex survey sampling, some definitions and the motivation of our research.

1.1 Recurrence risk definition and review of literature

Familial aggregation, also known as family aggregation, is the clustering of certain traits, behaviours, or disease among genetically related family members. Determining the extent of familial aggregation of a disease is important for identifying a potential genetic cause of the disease[7]. Studies showing familial aggregation provide the first step to identify the genes, which can lead to disease treatment, detection and prevention.

It is important to keep in mind that familial aggregation may arise because of either genetic or environmental similarities[7]. Even though genetic distributions are stable over time, when familial aggregation changes over time, this usually indicates changes in the enviroment related to the disease on trait of interest.

Recurrence risk is a measure of familial aggregation. Estimating recurrence risk requires the disease status about two or more family members where at least one of them is affected. A definition of recurrence risk as defined by the NCI dictionary of genetic terms (https://www.cancer.gov/publications/dictionaries/genetics-dictionary?cdrid=460213) is given as: "In genetics, recurrence risk is the likelihood that a hereditary trait or disorder present in one family member will occur again in other family members. It can be calculated as the probability that a family member of an affected individual is also affected." Since recurrence risk is a conditional probability, this probability can depend on which member's individual's disease status is being condi-

tioned, e.g., if the family members are mother and child, the relationship between the two family members is non-exchangeable, then whose disease status is conditioned on may matter; while if the relationship is exchangeable, e.g., siblings, then whose disease status is conditioned on will not matter.

In early studies of sibling recurrence risk estimation, ascertainment bias was a potential problem when affected individuals were not randomly sampled[16]. One approach to avoid ascertainment bias is to select a simple random sample of sibships from the population of sibships with at least one affected in each sibship and then estimate siblings recurrence risk[35]. This is called the complete ascertainment approach. Another approach is to randomly sample affected individuals from the population and then ascertain their sibships. The sample of sibships obtained are selected proportional to number of affected individuals in the sibships[15]. This approach is called the single ascertainment approach. Olson and Cordell[35] proposed consistant estimators of sibling recurrence risk and Zou and Zhao[47] provided variance estimators of sibling recurrence risk estiamte. However, both complete ascertainment approach and single ascertainment approach have problems. For complete ascertainment approach, population registries or sample frames of sibships from which simple random samples of sibships can be selected are not available except for specialized population such as for Mormons. For the single ascertainment approach, population registries from which to obtain simple random samples of affected individuals are limited. Thus, an alternative approach was provided by Graubard and Sirken[15] that extended the single ascertainment approach by using large population-based household surveys to report individuals' disease status and their siblings' disease status from random samples. This approach is called network sampling. We will use network sampling approach in our research to estimate recurrence risk.

Besides the recurrence risk, recurrence risk ratio is another measurement of familial aggregation. It is defined as the ratio by recurrence risk and the prevalence of disease.

The quadratic exponential model (QEM) has been used to estimate recurrence risk. The QEM was first introduced by Zhao and Prentice (1990) for a single binary outcome for each family member and was extended by Betensky and Whittemore (1996) for multiple binary ourcomes for each family member, and later Rabbee and Betensky (2004) derived power to perform sample size calculations for the QEM[31]. However, the QEM has the drawback of being irreproducible, in that the regression parameters depend on the QEM family size. For samples with varying family sizes, the interpretation of the regression is problematic. Thus, Matthews, Finkelstein and Betensky[31] introduced a marginal approach to model familial aggregation of varying family sizes. In their marginal approach, the joint distribution of the binary trait on disease among the family members is determined by a referent family size that is the smallest family size in the sample larger than one as familal aggregation cannot be computed from families with one member. In this approach, each family with size greater than the referent family size is subdivided into pseudo families with referent size by all possible combinations of family members of the referent family size. These pseudo families are combined across all families in the sample using QEM model to form a composite likelihood.

1.2 Background of complex survey sampling

A sample survey is "a study that obtains data from a subset of a population, in order to estimate population attributes" (http://stattrek.com/statistics/dictionary.aspx?definition=Sample+survey). Random samples that don't meet the assumptions of simple random samples are called complex samples (the definition was given by

the website: http://www.theanalysisfactor.com/complex-sample-what-and-why). In complex samples, the members of the finite target population often do not have equal probability of being included in the sample, i.e., self-weight sample design. In practice, household surveys use complex sample designs involving multistage stratified geographically clustered sampling. The sampling is often unequal probability sampling to oversample subgroups of the population such as people with low incomes. For example, in our real example in Chapter 5 that uses a national survey, the National Health Interviewed Survey (NHIS), which collected diabetes information in 1976, the complex survey design had 149 sampling strata and sampled 298 primary sampling units (PSU) at the first stage of sampling. Two further stages of sampling were conducted to sample participating households. We use the marginal QEM accounting for the stratified and clustered sampling along with differential probabilities of sampling in our method to study the familial aggregation.

The relevant details about the complex survey sampling and simple random sampling are outlined in the following parts.

- Simple random sampling (SRS): SRS is a basic sampling method. For SRS with a sample size of n, all possible subsets of size n individuals in a population of N individuals are equally likely to be selected[23]. There are two kinds of SRS: SRS with replacement (SRSWR) and SRS without replacement (SRSWOR), distinguished by whether each individual can be selected more than once. For SRSWOR, any individual cannot be chosen more than once. There are $\binom{N}{n}$ possible subsets in total, hence the probability for one subset is selected is $\frac{1}{\binom{N}{n}}$, while for i^{th} individual in the population, there are $\binom{N-1}{n-1}$ possible subsets which include individual i. Therefore, the probability that i^{th} individual is selected in the sample equals to $\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$ for any individual in the population. For

SRSWR, the same individual can be chosen more than once, hence there are N^n possible subsets in total, and the probability for one subset is selected is $\frac{1}{N^n}$. As for the possible ways that i^{th} individual is in the sample, it is a summation since i^{th} individual can be selected once, twice,..., n times at most. Hence, for the probability that i^{th} individual is seleted, we can use complement concept so that the probability equals to $1 - \frac{(N-1)^n}{N^n}$. There is one important concept that is often used in survey sampling: inclusion probability. Inclusion probability is the chance that i^{th} individual is selected in the sample. Therefore, for SRSWOR, inclusion probability is $\frac{n}{N}$ for any individual; while for SRSWR, inclusion probability is $1 - \frac{(N-1)^n}{N^n}$ for any individual. For example, suppose our population includes A, B, C, D, four elements, and we select two elements. Under SRSWOR, the inclusion probability that A is in the sample equals to $\frac{\binom{3}{1}}{\binom{4}{2}} = \frac{1}{2}$ (i.e., $\{A, B\}$, $\{A, C\}$ and $\{A, D\}$, 3 subsets out of all possible 6 subsets). Under SRSWR, the inclusion probability that A is in the sample is $1 - \frac{3^2}{4^2} = \frac{7}{16}$ (i.e., $\{A, A\}$, $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{B, A\}$, $\{C, A\}$, $\{D, A\}$, 7 subsets out of all possible 16 subsets). SRS is rarely used by itself in human population surveys. However, SRS is important since it can be a component of complex sampling designs.

- Stratified Random Sampling: Divide the whole population into several relatively homogeneous (with respect to distribution of important variables) and non-overlapping subgroups called strata, and then independently randomly select samples from each stratum. This sampling approach often reduces cost and time, and also increases precision over SRS with the same size. Stratified random sampling is useful especially when the measurement within the same stratum has smaller variance than the variance for the entire population. Also, different sampling methods can be used for different strata.

- Clustering Sampling: In clustering sampling, the population is divided into non-overlapping groups called primary sampling units, commonly abbreviated as PSU, and a random sample of PSUs are selected as the first stage sampling. At the next stages of sampling, units are randomly sampled from the sampled PSUs. Note that the individuals within the same PSU can be more similar than the individuals between PSUs. With respect to variables that are measured, this causes intracluster correlation.

- Network Sampling: "Conventional sampling and network sampling differ with regard to the number of different selection units at which the same population element is countable in the survey"[41]. In conventional sampling, every population element is linked to one and only one selection unit, e.g., sampled household, in the survey. In network sampling, each population element is a part of a network of selection units at which it is countable and the network size may vary for each population element[41]. Note that network size for a population element is the number of possible selection units that can report the population element. For example, in 1976 NHIS diabetes data, network sampling was used. The population elements were all siblings (alive or dead) of living people in the US in 1976. Each siblings in the population can be reported by a living sibling (selection unit). The network size for a reported sibling is the number of their living siblings in the US in 1976. Therefore, there are two kinds of weights: one is network weight which equals to the reciprocal of network size, and the other is sample weight which equals to the reciprocal of inclusion probability.

In the simulation studies in Chapter 6, informative sampling is considered in the sense that the selection probabilities are correlated with the disease trait. Sampling weights, which are the inverse of the selection probabilities, are assigned to each sampled individual. All the estimation and inference that we consider will involve weighting the observations by the sampled individual's sample weight. This weighting is needed to assure unbiased estimation and accurate hypothesis testing.

In surveys, counting rules specify conditions for linking population elements to selection units at which the population elements are eligible to be reported in the survey[41]. There are two counting rules which are usually considered in sibling recurrence risk estimation: All Sibling counting rule (AllSCR) where any individual who is selected in the sample reports about his/her siblings' disease status; Affected Sibling counting rule (AffSCR) where only affected individuals in the sample report disease status of his/her siblings. In diabetes studies, affected individuals are thought to report more accurately about their siblings' disease status than unaffected individuals[4]. In 1976 NHIS diabetes data we use, there is not only reporting about diabetes status for siblings but also for parents. Therefore, we will extend AllSCR, which means any individual who is selected in the sample can report his/her siblings' and parents' disease status. In other words, every sibling's disease status can be reported by his/her alive siblings and every parent's disease status can be reported by his/her alive children. In our 1976 NHIS diabetes study with network sampling approach, network size for AllSCR is the number of living siblings in the family. We will investigate estimation of the prevalence of disease among only living family members and also among both alive and dead family members, and compare recurrence risk estimates between these two conditions.

1.3 Some definitions

1.3.1 Quadratic exponential model

In recent years, many methods have been proposed for familial aggregation studies. One popular method among these is using the quadratic exponential model (QEM)[31]. We use QEM to specify the joint probability distribution of the disease status among the relatives in a family. This joint distribution is given by,

$$P(y_1, y_2, ..., y_n) = \Delta_n^{-1} exp(\delta \sum_{i=1}^{n} y_i + \gamma \sum_{i<j} y_i y_j), \tag{1}$$

where Δ_n is a normalization constant. y_i denotes the disease status of the i^{th} individual in a family and n is number of family members[31]. If the family size is 2, then we have the bivariate density:

$$P(y_1, y_2) = \frac{e^{\delta(y_1+y_2)+\gamma y_1 y_2}}{1 + 2e^{\delta} + e^{2\delta+\gamma}}.$$

In the QEM, the third and higher order associations are taken to be zero. In Equation 1, we assume all the disease status and pairwise of disease status are exchangeable, which means the probability of each family member's disease status given the disease status of another family member does not depend on the family relationship (e.g., parent-child or sib-sib). The parameters δ and γ have conditional interpretations for probability of disease status and for association.

$$exp(\delta) = \frac{P(y_i = 1 | y_{-i} = 0)}{P(y_i = 0 | y_{-i} = 0)},$$

$$exp(\gamma) = \frac{P(y_i = 1|y_j = 1, y_{-ij})/P(y_i = 0|y_j = 1, y_{-ij})}{P(y_i = 1|y_j = 0, y_{-ij})/P(y_i = 0|y_j = 0, y_{-ij})}$$
$$= \frac{P(y_i = 1, y_j = 1|y_{-ij})P(y_i = 0, y_j = 0|y_{-ij})}{P(y_i = 1, y_j = 0|y_{-ij})P(y_i = 0, y_j = 1|y_{-ij})},$$

where y_{-i} denotes the disease status within the family excluding the i^{th} individual; y_{-ij} denotes the outcomes within the family excluding the i^{th} and j^{th} individuals. It follows that δ is the log-odds which measures the probability of one family member is affected, given the others are not affected; while γ is the log-odds ratio, which represents the association between two family members[31].

However, as indicated before, the QEM only can be used when all the families share the same family size, which is unrealistic in the family disease studies. Several other approaches which are suitable for varying family size have been proposed. Among these approaches, the marginal approach is most useful because of its simplicity, unbiased parameter estimation and efficiency[31].

1.3.2 Marginal approach

Matthews et al.[31] proposed the marginal approach, "assume that smallest families are of size n and all subsets family members of size n from families of sizes that are larger than n follow the QEM". In other words, select the smallest family size as the referent family size. In the marginal approach, the score functions U are given by,

$$U = \begin{cases} T - E_{\delta,\gamma}(T) & \text{if } m = n \\ \sum_{g=1}^{G} T_g - \binom{m}{n} E_{\delta,\gamma}(T) & \text{if } m > n. \end{cases} \quad (2)$$

Here, $T = (\sum_{i=1}^{n} y_i, \sum_{i<j} y_i y_j)^T$, T_g is the statistics for the g^{th} subset and $G = \binom{m}{n}$ (Matthews, Finkelstein and Betensky, 2005).

In the marginal approach, we initially assume the subsets arising from the same family are independent with each other, even though they are actually dependent because the same family members are in multiple subsets. To adjust for this dependence, robust variance estimation is used, and is calculated via a Taylor series expansion of the sum of the family-specific scores. The variance-covariance matrix of the parameter estimates is given by,

$$Cov(\hat{\Phi}) = (\sum_{i=1}^{n} \frac{\partial U_i(\Phi)}{\partial \Phi})^{-1} Cov(\sum_{i=1}^{n} U_i(\Phi))((\sum_{i=1}^{n} \frac{\partial U_i(\Phi)}{\partial \Phi})^{-1})^T,$$

where i is the index of family[31]. To estimate the variance-covariance matrix, we use

$$\hat{cov}(\hat{\Phi}) = (\sum_{i=1}^{n} \frac{\partial U_i(\Phi)}{\partial \Phi}|_{\Phi=\hat{\Phi}})^{-1} \hat{cov}(\sum_{i=1}^{n} U_i(\Phi))((\sum_{i=1}^{n} \frac{\partial U_i(\Phi)}{\partial \Phi}|_{\Phi=\hat{\Phi}})^{-1})^T. \quad (3)$$

Usually, $\sum_{i=1}^{n} U_i(\hat{\Phi}) U_i(\hat{\Phi})^T$ is used to estimate $Cov(\sum_{i=1}^{n} U_i(\Phi))$. But in our study with complex survey sampling, the estimator of the middle part $Cov(\sum_{i=1}^{n} U_i(\Phi))$ is derived based on the complex sample design of the survey.

1.3.3 Composite likelihood

The parameter estimates are obtained after setting the score function to be zero and solving the score equations for the parameters, but how can we determine the score equations? We derive the score equations from a composite likelihood. Composite likehood is an inference function derived by multiplying a collection of component likelihoods. Because the component likelihoods are based on marginal or conditional densities, the estimating equation derived from the composite log-likehoood

is unbiased[45]. Hence, unbiased estimation of parameters can be obtained by setting the estimating function to be zero. This is accomplished without specifying the actual joint distribution. There are many different applications of composite likelihoods, such as analysis of longitudinal data, time series, statistical genetic data, etc. Varin and et al.[45] provide a comprehensive review of composite likelihood methods.

The original form of composite likelihood is a weighted product, proposed by Lindsay (1988),

$$L_C(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k},$$

where the $w_k,\ k = 1, ..., K$ are nonnegative weights[45]. Note that if all the weights w_k are the same, then they can be ignored. The composite likelihood can be also called a quasi-likelihood since it is not a real likelihood function.

There are many other variations of composite likelihood derived from marginal or conditional lower dimensional densities. In our research, we will use the version from Cox and Reid[9], the pairwise form is given by,

$$L_C(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \theta).$$

That is, given the pairwise densities, we can derive the composite likelihood for higher dimensional distribution even though we do not specify the full joint distribution.

Generalized estimating equation (GEE), was initially introduced by Liang and

Zeger in 1986[29]. GEE is used to get the parameter estimates of a generalized linear model with possible unknown correlated responses. The model does not need to satisfy homogeneity of variances or independence of reponses. The quasi-likelhood estimation rather than maximum likelihood estimation (MLE), which requires correct specification of the joint distributions, is used to obtain the parameter estimates. We emphasize that MLE is difficult to obtain because it is difficult, without making strong distributional assumptions, to specify the likelihood. The estimation equations can be obtained from the composite likelihood as mentioned in previous paragraph.

1.4 Thesis objective

Most research on familial aggregation uses sibling recurrence risk and does not consider other types of family relationships. It is commonly believed that family relationship has influence on familial aggregation because of genetics. Thus, in our research, our purpose is to study the relationship between recurrence risk of disease and family relationship using 1976 NHIS diabetes data. We will propose a relationship-EQEM. Our another purpose is to study the relationship between sibling recurrence risk and family-level covariates (e.g., race) and we refer to this as race-EQEM. Even though race-EQEM is a model-based approach, we will show equivalency between the estimation results from our model and estimation results from nonparametric approach proposed by Graubard and Sirken in 2013 using 1976 NHIS diabetes data[15]. We will extend the parameter estimation and variance estimation along with hypothesis testing of model parameters and recurrence risk to apply to data from complex samples such as stratified multistage cluster sampling used in household surveys like the NHIS. Our approach can be applied for varying family sizes utilizing marginal estimation, and it is easy to interpret results and conduct the computation as it is an extension of the QEM.

The following is the outline of the thesis: In Chapter 2, our extended new models with different covariates will be proposed (i.e., relationship-EQEM and race-EQEM). In Chapter 3, assorted complex sampling methods are incoporated in our proposed models including adding sample weights to sample units using specific sample designs, for example, stratified sampling, clustering sampling, and network sampling. The estimation procedures for parameters and variance estimators will be adapted to these complex survey design methods. Also, propensity score analysis will introduce to adjust for individual-level covariates such as age. In Chapter 4, we will conduct hypothesis testing with complex survey data using Wald tests and Quasi-score tests[37]. In Chapters 5 and 6, simulation studies and real data analysis are used to investigate finite sample properties of our estimation and hypothesis testing methods. For real data analysis, we use the diabetes data in 1976 NHIS. In Chapter 7, we provide a summary and discussion of our research findings and limitation, and areas of future researches.

Chapter 2 Extension of QEM with covariates

We use the marginal approach to analyze the familial aggregation with varying family sizes. "Family" refers to the specific relationships that are of interest for determining aggregation of a trait. For example, the relationship could be siblings. In the marginal approach, the smallest family size greater or equal to two is said to be the referent family size. The referent family size is usually 2, hence the referent density of the family members' disease status is usually a bivariate density.

The purpose of our model is to accout for covariates in the bivariate density, in order to extend the model application, such as covariates for pairs of family members (e.g., family relationship), family-level covariate (e.g., race), or individual-level covariate (e.g., age). In the following sections, different extended QEMs (EQEM) are considered.

2.1 Pairwise-level covariate

Family relationship is an important covariate among pairs of family members. For members that are siblings, the disease status can be exchangeable (i.e., ordering of the individuals in conditional probabilities is unnecessary, $P(y_1 = 1|y_2 = 1) = P(y_2 = 1|y_1 = 1)$) if individual-level variables are not considered, since all the siblings have the same relationship (half-siblings are not considered), however, after incorporating different family relationships, such as parent-child relationships, into the same model, we may loose exchangeability.

When there are different relationships between family members, the type of disease or trait under study may dictate whether exchangeability can be assumed. For example, if the families consist of parents and children and disease is the flu, then

exchangeability could be a reasonable assumption. However, for a disease such as diabetes, which is genetically based, exchangeability would not be reasonable. Therefore, we propose two different approaches: Approach 1 is no ordering accounted for among family members' disease status, and Approach 2 considers ordering.

We will mainly focus on familial aggregation that is thought to be genetically based in the subsequent work. Thus, we do not consider the Father-Mother relationship. Also, to simplify derivations, calculations for families of size 2 and 3 are shown in detail, these derivations can be extended to larger families. Moreover, we restrict ourselves to three types pairwise family relationships: Father-Child, Mother-Child, and Child-Child here. However, with extended family data, other relationships such as ancle-nephew or grandparent-grandchild are possible and can be handled similarly to parent-child/child-child relationships.

The followings are the extended bivariate densities with different family relationships:

Approach 1:

$$P(y_1, y_2|I_1, I_2) = \frac{e^{(\delta+\beta_1 I_1+\beta_2 I_2)(y_1+y_2)+(\gamma+\beta_3 I_1+\beta_4 I_2)y_1 y_2}}{1+2e^{\delta+\beta_1 I_1+\beta_2 I_2}+e^{2\delta+2\beta_1 I_1+2\beta_2 I_2+\gamma+\beta_3 I_1+\beta_4 I_2}};$$

Approach 2:

$$P(y_1, y_2|I_1, I_2) = \frac{e^{(\delta+\beta_1 I_1+\beta_2 I_2)y_1+(\delta+\beta_3 I_1+\beta_4 I_2)y_2+(\gamma+\beta_5 I_1+\beta_6 I_2)y_1 y_2}}{1+e^{\delta+\beta_1 I_1+\beta_2 I_2}+e^{\delta+\beta_3 I_1+\beta_4 I_2}+e^{2\delta+\beta_1 I_1+\beta_2 I_2+\gamma+\beta_3 I_1+\beta_4 I_2+\beta_5 I_1+\beta_6 I_2}}, \tag{4}$$

where $I_1 = 1, I_2 = 0$, if Father-Child; $I_1 = 0, I_2 = 1$, if Mother-Child; $I_1 = 0, I_2 = 0$, if

Child-Child. Therefore, there are six parameters δ, γ, β_1, β_2, β_3, and β_4 in Approach 1. There are eight parameters δ, γ, β_1, β_2, β_3, β_4, β_5 and β_6 in Approach 2 because ordering needs to be taken into account. We call Approach 2 "relationship-EQEM".

It is likely that the ordering is important in the NHIS diabetes data analysis, so Approach 2 is what we used for following research, even though Approach 1 is also useful in familial aggregation studies. For the bivariate density, let y_1 denote the disease status of parent for Father-Child and Mother-Child relationships and y_2 denote the disease status of child; while if it is Child-Child relationship, then y_1 and y_2 are exchangeable.

2.1.1 Parameter estimation

Father-Child: For one family of size 2 with Father-Child relationship, the likelihood function is given by,

$$L^{(FC)} = \frac{e^{(\delta+\beta_1)y_1+(\delta+\beta_3)y_2+(\gamma+\beta_5)y_1y_2}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},$$

where y_1 denotes the disease status of father and y_2 denotes the disease status of child. We form the log-likelihood and take derivatives with respect to parameters δ, γ, β_1, β_3 and β_5 to obtain the score functions,

$$
\begin{aligned}
\frac{\partial l^{(FC)}}{\partial \delta} &= y_1 + y_2 - \frac{e^{\delta+\beta_1} + e^{\delta+\beta_3} + 2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}, \\
\frac{\partial l^{(FC)}}{\partial \gamma} &= y_1 y_2 - \frac{e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}, \\
\frac{\partial l^{(FC)}}{\partial \beta_1} &= y_1 - \frac{e^{\delta+\beta_1} + 2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}, \\
\frac{\partial l^{(FC)}}{\partial \beta_3} &= y_2 - \frac{e^{\delta+\beta_3} + 2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}, \\
\frac{\partial l^{(FC)}}{\partial \beta_5} &= y_1 y_2 - \frac{e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}.
\end{aligned}
$$

Mother-Child: For one family of size 2 with Mother-Child relationship, the likelihood function is given by,

$$
L^{(MC)} = \frac{e^{(\delta+\beta_2)y_1 + (\delta+\beta_4)y_2 + (\gamma+\beta_6)y_1 y_2}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}},
$$

where y_1 denotes the disease status of mother and y_2 denotes the disease status of child. Then, the score functions are obtained as follows,

$$
\begin{aligned}
\frac{\partial l^{(MC)}}{\partial \delta} &= y_1 + y_2 - \frac{e^{\delta+\beta_2} + e^{\delta+\beta_4} + 2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}, \\
\frac{\partial l^{(MC)}}{\partial \gamma} &= y_1 y_2 - \frac{e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}, \\
\frac{\partial l^{(MC)}}{\partial \beta_2} &= y_1 - \frac{e^{\delta+\beta_2} + 2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}, \\
\frac{\partial l^{(MC)}}{\partial \beta_4} &= y_2 - \frac{e^{\delta+\beta_4} + 2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}, \\
\frac{\partial l^{(MC)}}{\partial \beta_6} &= y_1 y_2 - \frac{e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}.
\end{aligned}
$$

Child-Child: For one family of size 2 with Child-Child relationship, the likelihood function is given by,

$$L^{(CC)} = \frac{e^{\delta(y_1+y_2)+\gamma y_1 y_2}}{1+2e^{\delta}+e^{2\delta+\gamma}}.$$

The score functions are derived as,

$$\begin{aligned}
\frac{\partial l^{(CC)}}{\partial \delta} &= y_1 + y_2 - \frac{2e^{\delta} + 2e^{2\delta+\gamma}}{1 + 2e^{\delta} + e^{2\delta+\gamma}}, \\
\frac{\partial l^{(CC)}}{\partial \gamma} &= y_1 y_2 - \frac{e^{2\delta+\gamma}}{1 + 2e^{\delta} + e^{2\delta+\gamma}}.
\end{aligned}$$

Father-Child-Child: For one family of size 3 with Father-Child-Child relationship, we subgroup the family into three pairs of family members: Father-Child, Father-Child and Child-Child. That is, we treat the family of size 3 with Father-Child-Child relationship as three sub-families of size 2 with Father-Child, Father-Child and Child-Child relationships. Note that y_1 denotes the disease status of father, y_2 and y_3 denote the disease status of children.

Mother-Child-Child: Similarly, for one family of size 3 with Mother-Child-Child relationship, we subgroup the family into three sub-families of size 2: Mother-Child, Mother-Child and Child-Child. Note that y_1 denotes the disease status of mother, y_2 and y_3 denote the disease status of children.

Child-Child-Child: For one family of size 3 with Child-Child-Child relationship,

we subgroup the family into three sub-families of size 2 with Child-Child relationships.

Father-Mother-Child: For one family of size 3 with Father-Mother-Child relationship, we subgroup the family into two sub-families of size 2: Father-Child and Mother-Child. Note that Father-Mother relationship is not considered because they are not typically genetically related. Note that y_1 denotes the disease status of father, y_2 represents the disease status of mother and y_3 denotes the disease status of child.

Given the pairwise likelihood, a composite likelihood function can be characterized by multiplying the pairwise likelihoods[45]. Hence, we can write down a composite likelihood with different relationships for families of larger sizes based on the bivariate density we have. Suppose there are N_1, N_2, N_3 families of size 2 and relationship Father-Child, Mother-Child, Child-Child, respectively. There are N_4, N_5, N_6, N_7 families of size 3 and relationship Father-Child-Child, Mother-Child-Child, Child-Child-Child, Father-Mother-Child, respectively. Then, the composite likelihood function and the score functions for all the families is as follows,

$$
\begin{aligned}
L^{(total)} &= (N_1 + 2N_4 + N_7)L^{(FC)} \times (N_2 + 2N_5 + N_7)L^{(MC)} \times (N_3 + N_4 + N_5 + 3N_6)L^{(CC)} \\
&= (N_1 + 2N_4 + N_7)\frac{e^{(\delta+\beta_1)y_1+(\delta+\beta_3)y_2+(\gamma+\beta_5)y_1y_2}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}} \\
&\times (N_2 + 2N_5 + N_7)\frac{e^{(\delta+\beta_2)y_1+(\delta+\beta_4)y_2+(\gamma+\beta_6)y_1y_2}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}} \\
&\times (N_3 + N_4 + N_5 + 3N_6)\frac{e^{\delta(y_1+y_2)+\gamma y_1y_2}}{1 + 2e^{\delta} + e^{2\delta+\gamma}},
\end{aligned}
$$

$$\frac{\partial l^{(total)}}{\partial \delta} = \sum_{i=1}^{N_1+N_2+N_3}(y_{i1}+y_{i2}) + 2\sum_{i=1}^{N_4+N_5+N_6}(y_{i1}+y_{i2}+y_{i3}) + \sum_{i=1}^{N_7}(y_{i1}+y_{i3}+y_{i2}+y_{i3})$$
$$- (N_1+2N_4+N_7)\frac{e^{\delta+\beta_1}+e^{\delta+\beta_3}+2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}$$
$$- (N_2+2N_5+N_7)\frac{e^{\delta+\beta_2}+e^{\delta+\beta_4}+2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1+e^{\delta+\beta_2}+e^{\delta+\beta_4}+e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}$$
$$- (N_3+N_4+N_5+3N_6)\frac{2e^{\delta}+2e^{2\delta+\gamma}}{1+2e^{\delta}+e^{2\delta+\gamma}},$$

$$\frac{\partial l^{(total)}}{\partial \gamma} = \sum_{i=1}^{N_1+N_2+N_3} y_{i1}y_{i2} + \sum_{i=1}^{N_4+N_5+N_6}(y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3}) + \sum_{i=1}^{N_7}(y_{i1}y_{i3}+y_{i2}y_{i3})$$
$$- (N_1+2N_4+N_7)\frac{e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}$$
$$- (N_2+2N_5+N_7)\frac{e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1+e^{\delta+\beta_2}+e^{\delta+\beta_4}+e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}$$
$$- (N_3+N_4+N_5+3N_6)\frac{e^{2\delta+\gamma}}{1+2e^{\delta}+e^{2\delta+\gamma}},$$

$$\frac{\partial l^{(total)}}{\partial \beta_1} = \sum_{i=1}^{N_1} y_{i1} + 2\sum_{i=1}^{N_4} y_{i1} + \sum_{i=1}^{N_7} y_{i1}$$
$$- (N_1+2N_4+N_7)\frac{e^{\delta+\beta_1}+2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},$$

$$\frac{\partial l^{(total)}}{\partial \beta_2} = \sum_{i=1}^{N_2} y_{i1} + 2\sum_{i=1}^{N_5} y_{i1} + \sum_{i=1}^{N_7} y_{i2}$$
$$- (N_2+2N_5+N_7)\frac{e^{\delta+\beta_2}+2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1+e^{\delta+\beta_2}+e^{\delta+\beta_4}+e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}},$$

$$\frac{\partial l^{(total)}}{\partial \beta_3} = \sum_{i=1}^{N_1} y_{i2} + 2\sum_{i=1}^{N_4} y_{i2} + \sum_{i=1}^{N_7} y_{i3} - (N_1 + 2N_4 + N_7)\frac{e^{\delta+\beta_3} + 2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},$$

$$\frac{\partial l^{(total)}}{\partial \beta_4} = \sum_{i=1}^{N_2} y_{i2} + 2\sum_{i=1}^{N_5} y_{i2} + \sum_{i=1}^{N_7} y_{i3} - (N_2 + 2N_5 + N_7)\frac{e^{\delta+\beta_4} + 2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}},$$

$$\frac{\partial l^{(total)}}{\partial \beta_5} = \sum_{i=1}^{N_1} y_{i1}y_{i2} + \sum_{i=1}^{N_4}(y_{i1}y_{i2} + y_{i1}y_{i3}) + \sum_{i=1}^{N_7} y_{i1}y_{i3} - (N_1 + 2N_4 + N_7)\frac{e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},$$

$$\frac{\partial l^{(total)}}{\partial \beta_6} = \sum_{i=1}^{N_2} y_{i1}y_{i2} + \sum_{i=1}^{N_5}(y_{i1}y_{i2} + y_{i1}y_{i3}) + \sum_{i=1}^{N_7} y_{i2}y_{i3} - (N_2 + 2N_5 + N_7)\frac{e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}},$$

where $l^{(total)} = logL^{(total)}$.

Finally, all the score functions are set equal to zero to solve for the parameter estimates $\hat{\delta}$, $\hat{\gamma}$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$ and $\hat{\beta}_6$. We will discuss the variance estimation in next chapter.

2.1.2 Recurrence risk estimation

If Approach 2 is issued where ordering is accounted for, then the estimated probability that y_1 is affected given y_2 is affected ($RR_{1|2}$) and the estimated propbability that y_2 is affected given y_1 is affected ($RR_{2|1}$) can differ. After obtaining the parameter

estimates $\hat{\delta}$, $\hat{\gamma}$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$ and $\hat{\beta}_6$, the estimated recurrence risks are given by,

$$\begin{aligned}
\hat{RR}_{2|1} &= \hat{P}(y_2 = 1|y_1 = 1) = \frac{\hat{P}(y_1 = 1, y_2 = 1)}{\hat{P}(y_1 = 1)} = \frac{\hat{P}(y_1 = y_2 = 1)}{\hat{P}(y_1 = 1, y_2 = 0) + \hat{P}(y_1 = y_2 = 1)} \\
&= \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_3 I_1+\hat{\beta}_4 I_2+\hat{\beta}_5 I_1+\hat{\beta}_6 I_2}}{1 + e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_3 I_1+\hat{\beta}_4 I_2+\hat{\beta}_5 I_1+\hat{\beta}_6 I_2}} \\
&= \begin{cases} \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_3+\hat{\beta}_5}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_3+\hat{\beta}_5}} & \text{if Father-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_4+\hat{\beta}_6}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_4+\hat{\beta}_6}} & \text{if Mother-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}}}{1+e^{\hat{\delta}+\hat{\gamma}}} & \text{if Child-Child,} \end{cases}
\end{aligned} \tag{5}$$

$$\begin{aligned}
\hat{RR}_{1|2} &= \hat{P}(y_1 = 1|y_2 = 1) = \frac{\hat{P}(y_1 = 1, y_2 = 1)}{\hat{P}(y_2 = 1)} = \frac{\hat{P}(y_1 = y_2 = 1)}{\hat{P}(y_1 = 0, y_2 = 1) + \hat{P}(y_1 = y_2 = 1)} \\
&= \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_1 I_1+\hat{\beta}_2 I_2+\hat{\beta}_5 I_1+\hat{\beta}_6 I_2}}{1 + e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_1 I_1+\hat{\beta}_2 I_2+\hat{\beta}_5 I_1+\hat{\beta}_6 I_2}} \\
&= \begin{cases} \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_1+\hat{\beta}_5}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_1+\hat{\beta}_5}} & \text{if Father-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_2+\hat{\beta}_6}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_2+\hat{\beta}_6}} & \text{if Mother-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}}}{1+e^{\hat{\delta}+\hat{\gamma}}} & \text{if Child-Child.} \end{cases}
\end{aligned} \tag{6}$$

From the $\hat{RR}_{2|1}$ and $\hat{RR}_{1|2}$ under our model formulation, we can see that the estimated recurrence risk for two siblings is invariant to the ordering, that is, $\hat{RR}_{2|1} = \hat{RR}_{1|2}$ for Child-Child relationship.

2.2 Family-level covariate

In this section, one family-level covariate is added to the QEM, e.g., race. We assume that all family members share the same race within a family, but race can differ among different families. Hence, we allow the parameters in our models to differ between different races. For example, in the 1976 NHIS diabetes data, there are only

two possible races: white and other race, and we consider only sibling relationships in the model. Thus δ_W, γ_W; δ_O, γ_O are used to represent the parameters for the whites and other races respectively.

Thus, the bivariate density for different races is given by,

$$\begin{aligned} \textbf{Whites:} \qquad & P(y_1, y_2|\text{race=W}) = \frac{e^{\delta_W(y_1+y_2)+\gamma_W y_1 y_2}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}}, \\ \textbf{Other races:} \qquad & P(y_1, y_2|\text{race=O}) = \frac{e^{\delta_O(y_1+y_2)+\gamma_O y_1 y_2}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}}. \end{aligned} \tag{7}$$

We call the above model "race-EQEM".

2.2.1 Parameter estimation

For one single family with a particular race and sibling relationship, the composite likelihood functions and score functions are shown as follows (Note that the referrent family size is still 2, and we only study families of size 2 and 3 as an example),

Whites:

$$L_W = \frac{e^{\delta_W(y_1+y_2)+\gamma_W y_1 y_2}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}}.$$

Then, the score functions are derived as:

$$\begin{aligned} \frac{\partial l_W}{\partial \delta_W} &= y_1 + y_2 - \frac{2e^{\delta_W}+2e^{2\delta_W+\gamma_W}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}}, \\ \frac{\partial l_W}{\partial \gamma_W} &= y_1 y_2 - \frac{e^{2\delta_W+\gamma_W}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}}, \\ \frac{\partial l_W}{\partial \delta_O} &= \frac{\partial l_W}{\partial \gamma_O} = 0. \end{aligned}$$

Other races:

$$L_O = \frac{e^{\delta_O(y_1+y_2)+\gamma_O y_1 y_2}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}}.$$

Then, the score functions are derived as:

$$\begin{aligned}
&\frac{\partial l_O}{\partial \delta_W} = \frac{\partial l_O}{\partial \gamma_W} = 0,\\
&\frac{\partial l_O}{\partial \delta_O} = y_1 + y_2 - \frac{2e^{\delta_O}+2e^{2\delta_O+\gamma_O}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}},\\
&\frac{\partial l_O}{\partial \gamma_O} = y_1 y_2 - \frac{e^{2\delta_O+\gamma_O}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}}.
\end{aligned}$$

Suppose there are N_{W1} white families of size 2, N_{W2} white families of size 3; N_{O1} other race families of size 2, N_{O2} other race families of size 3. Then, the composite likelihood function and the score functions for all the families are presented by,

$$\begin{aligned}
L^{(total)} &= (N_{W1}+3N_{W2})L_W \times (N_{O1}+3N_{O2})L_O\\
&= (N_{W1}+3N_{W2})\frac{e^{\delta_W(y_1+y_2)+\gamma_W y_1 y_2}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}} \times (N_{O1}+3N_{O2})\frac{e^{\delta_O(y_1+y_2)+\gamma_O y_1 y_2}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}},
\end{aligned}$$

$$\frac{\partial l^{(total)}}{\partial \delta_W} = \sum_{i=1}^{N_{W1}}(y_{i1}+y_{i2}) + 2\sum_{i=1}^{N_{W2}}(y_{i1}+y_{i2}+y_{i3}) - (N_{W1}+3N_{W2})\frac{2e^{\delta_W}+2e^{2\delta_W+\gamma_W}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}},$$

$$\frac{\partial l^{(total)}}{\partial \gamma_W} = \sum_{i=1}^{N_{W1}} y_{i1}y_{i2} + \sum_{i=1}^{N_{W2}}(y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3}) - (N_{W1}+3N_{W2})\frac{e^{2\delta_W+\gamma_W}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}},$$

$$\frac{\partial l^{(total)}}{\partial \delta_O} = \sum_{i=1}^{N_{O1}}(y_{i1}+y_{i2}) + 2\sum_{i=1}^{N_{O2}}(y_{i1}+y_{i2}+y_{i3}) - (N_{O1}+3N_{O2})\frac{2e^{\delta_O}+2e^{2\delta_O+\gamma_O}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}},$$

$$\frac{\partial l^{(total)}}{\partial \gamma_O} = \sum_{i=1}^{N_{O1}} y_{i1}y_{i2} + \sum_{i=1}^{N_{O2}}(y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3}) - (N_{O1}+3N_{O2})\frac{e^{2\delta_O+\gamma_O}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}}.$$

Lastly, all the score functions are set to be zero to obtain the parameter estimates $\hat{\delta}_W$, $\hat{\gamma}_W$, $\hat{\delta}_O$ and $\hat{\gamma}_O$.

2.2.2 Recurrence risk estimation

Since only sibling relationships exist in the model, there is no consideration of ordering when calculating the recurrence risk, and y_1 and y_2 can be exchangeable at this time.

$$\hat{RR}_{race} = \hat{P}(y_2=1|y_1=1) = \begin{cases} \frac{e^{\hat{\delta}_W+\hat{\gamma}_W}}{1+e^{\hat{\delta}_W+\hat{\gamma}_W}} & \text{if whites} \\ \frac{e^{\hat{\delta}_O+\hat{\gamma}_O}}{1+e^{\hat{\delta}_O+\hat{\gamma}_O}} & \text{if other races.} \end{cases} \tag{8}$$

2.3 Individual-level covariate

As for a categotical individual-level covariate, such as gender, we can treat it as a pairwise-level covariate. Hence, there are Female-Female, Male-Male and Female-Male three different gender combinations and the model in Section 2.1 can be applied. However, this will not address individual-level covariates that are continuous. For example, age.

We initially tried to include a continuous covariate directly into the QEM. Suppose a_1 is the age of the first family member and a_2 is the age for the second one. Then the bivariate model we tried was,

$$P(y_1, y_2|a_1, a_2) = \frac{e^{\delta(y_1+y_2)+\gamma y_1 y_2+\gamma_1 a_1+\gamma_2 a_2}}{(1+2e^{\delta}+e^{2\delta+\gamma})e^{\gamma_1 a_1+\gamma_2 a_2}} = \frac{e^{\delta(y_1+y_2)+\gamma y_1 y_2}}{1+2e^{\delta}+e^{2\delta+\gamma}}.$$

As can be seen values of age in this expression have no effect on the density. Then, we considered transforming the continuous covariate into a categorical covariate. For instance, suppose there are three possible age combinations for pairs of siblings in our model: Young-Young (both of siblings are younger than 18 years old), Old-Old (both of siblings are older than 18 years old), Mixture (one sibling is older than 18 years old, while the other is younger than 18 years old), and we assign δ_Y, γ_Y; δ_O, γ_O; δ_M, γ_M to represent the parameters for the Young-Young pair, Old-Old pair and Mixture respectively. Here, age becomes a pairwise-level covariate. The disadvantage of using categorical age is that the adjustment for age will not be complete compared to adjust for age as a continuous variable.

Another approach was to adjust for age as a continuous variable by using propensity score weighting. The details are discussed in Chapter 3.

Chapter 3 Application of extended QEM (EQEM) with complex survey sampling

A major aim of this thesis is to develop inferential methods for recurrence risk that can be applied to complex survey data. This aim is motivated by the diabetes status data collected for sampled persons and his/her relatives in the 1976 NHIS. In Chapter 5, we will use the NHIS data to illustrate our methods. In this chapter, the EQEM is extended to complex survey data, which can have sample weights, and we propose robust variance estimation using the Taylor linearization method. At the end of this chapter, we show how propensity score weighting can be used to adjust for continuous covariates.

3.1 Parameter estimation with complex survey sampling

In Chapter 2, we presented the parameter estimation for simple random samples, where each sampled individual proband has the same probability of selection, therefore, requires no sample weighting. In complex samples, the probability of selection usually varies across sampled individuals and therfore their sample weights are not the same value. Therefore, estimation where observations are weighted by the samples is required in order to generally obtain approximately unbiased estimators.

In addition to the complex sampling of probands, the reporting of disease/trait information about relatives uses network sampling. Briefly, network sampling is when a sampled person can report the disease/trait for a defined network of units in the population (A description of network sampling was given in Section 1.2). In our motivating example of diabetes in the 1976 NHIS, the units are family members such as siblings and parents of the sampled person, and the counting rule (defines which units a sampled person can report) determines the network size. For example, if

the counting rule is that a sampled person/proband reports about their siblings and parents' disease/trait status, then the networks size is the number of living siblings plus the proband. Because larger families have higher probabilities of being reported in the survey under this counting rule, then the disease status of a reported family is downweighted by a network weight. We will use the inverse of network size as the network weight that is used for the downweighting. For example, if the network size is the number of living siblings plus the proband, s_i for i^{th} sampled proband, then the weight is $\frac{w_i}{s_i}$, where w_i is sample weight of proband.

Thus, the composite likelihood function and the score functions for "race-EQEM" are shown as follows (suppose n_{W1}, n_{W2}, n_{O1}, n_{O2}, $L_W^{(2)}$, $L_W^{(3)}$, $L_O^{(2)}$, $L_O^{(3)}$ denote the sample sizes and likelihoods of white families with number of living siblings of size 2, 3, other race families with number of living siblings of sizes 2, 3 respectively),

$$
\begin{aligned}
L^{(sample)} &= \sum_{i=1}^{n_{W1}} \frac{w_i}{2} L_W^{(2)} \times \sum_{i=1}^{n_{W2}} \frac{w_i}{3} L_W^{(3)} \times \sum_{i=1}^{n_{O1}} \frac{w_i}{2} L_O^{(2)} \times \sum_{i=1}^{n_{O2}} \frac{w_i}{3} L_O^{(3)} \\
&= \sum_{i=1}^{n_{W1}} \frac{w_i}{2} \frac{e^{\delta_W (y_{i1}+y_{i2}) + \gamma_W y_{i1} y_{i2}}}{1 + 2e^{\delta_W} + e^{2\delta_W + \gamma_W}} \times \sum_{i=1}^{n_{W2}} \frac{w_i}{3} \frac{e^{2\delta_W (y_{i1}+y_{i2}+y_{i3}) + \gamma_W (y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3})}}{(1 + 2e^{\delta_W} + e^{2\delta_W + \gamma_W})^3} \\
&\times \sum_{i=1}^{n_{O1}} \frac{w_i}{2} \frac{e^{\delta_O (y_{i1}+y_{i2}) + \gamma_O y_{i1} y_{i2}}}{1 + 2e^{\delta_O} + e^{2\delta_O + \gamma_O}} \times \sum_{i=1}^{n_{O2}} \frac{w_i}{3} \frac{e^{2\delta_O (y_{i1}+y_{i2}+y_{i3}) + \gamma_O (y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3})}}{(1 + 2e^{\delta_O} + e^{2\delta_O + \gamma_O})^3},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l^{(sample)}}{\partial \delta_W} &= \sum_{i=1}^{n_{W1}} \frac{w_i}{2} (y_{i1} + y_{i2}) + 2 \sum_{i=1}^{n_{W2}} \frac{w_i}{3} (y_{i1} + y_{i2} + y_{i3}) \\
&\quad - (\sum_{i=1}^{n_{W1}} \frac{w_i}{2} + 3 \sum_{i=1}^{n_{W2}} \frac{w_i}{3}) \frac{2e^{\delta_W} + 2e^{2\delta_W + \gamma_W}}{1 + 2e^{\delta_W} + e^{2\delta_W + \gamma_W}},
\end{aligned}
$$

$$\frac{\partial l^{(sample)}}{\partial \gamma_W} = \sum_{i=1}^{n_{W1}} \frac{w_i}{2} y_{i1}y_{i2} + \sum_{i=1}^{n_{W2}} \frac{w_i}{3}(y_{i1}y_{i2} + y_{i1}y_{i3} + y_{i2}y_{i3}) - (\sum_{i=1}^{n_{W1}} \frac{w_i}{2} + 3\sum_{i=1}^{n_{W2}} \frac{w_i}{3}) \frac{e^{2\delta_W+\gamma_W}}{1 + 2e^{\delta_W} + e^{2\delta_W+\gamma_W}},$$

$$\frac{\partial l^{(sample)}}{\partial \delta_O} = \sum_{i=1}^{n_{O1}} \frac{w_i}{2}(y_{i1} + y_{i2}) + 2\sum_{i=1}^{n_{O2}} \frac{w_i}{3}(y_{i1} + y_{i2} + y_{i3}) - (\sum_{i=1}^{n_{O1}} \frac{w_i}{2} + 3\sum_{i=1}^{n_{O2}} \frac{w_i}{3}) \frac{2e^{\delta_O} + 2e^{2\delta_O+\gamma_O}}{1 + 2e^{\delta_O} + e^{2\delta_O+\gamma_O}},$$

$$\frac{\partial l^{(sample)}}{\partial \gamma_O} = \sum_{i=1}^{n_{O1}} \frac{w_i}{2} y_{i1}y_{i2} + \sum_{i=1}^{n_{O2}} \frac{w_i}{3}(y_{i1}y_{i2} + y_{i1}y_{i3} + y_{i2}y_{i3}) - (\sum_{i=1}^{n_{O1}} \frac{w_i}{2} + 3\sum_{i=1}^{n_{O2}} \frac{w_i}{3}) \frac{e^{2\delta_O+\gamma_O}}{1 + 2e^{\delta_O} + e^{2\delta_O+\gamma_O}}.$$

The score functions are set to be zero in order to solve for the parameter estimates $\hat{\delta}_W$, $\hat{\gamma}_W$, $\hat{\delta}_O$ and $\hat{\gamma}_O$ based on the samples.

As for "relationship-EQEM", since each proband can report his/her parents' and siblings' disease information, the network size is also the number of living siblings for each family in the sample.

3.2 Recurrence risk, prevalence and recurrence risk ratio calculations with complex survey sampling

3.2.1 Recurrence risk calculation with complex survey sampling

Recurrence risk can be estimated by plugging the parameter estimates into $\hat{RR}_{2|1}$ (See Equation 5[1]), $\hat{RR}_{1|2}$ (See Equation 6[2]) or $\hat{RR}_{race}$ (See Equation 8[3]). If it is only sibling recurrence risk, we can also apply the method of Graubard and Sirken (2013)[15] for recurrence risk calculation with complex survey sampling. In that paper, sibling recurrence risk can be expressed as the proportion of affected siblings among all siblings of affected individuals in a population, that is,

$$\hat{SRR}_G = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i(a_i - 1)}{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i(s_i - 1)}, \tag{9}$$

where n is the sample size, and a_i is the number of affected siblings in sibship i. Note that if $a_i = 0$ or $s_i = 1$ for all the sibships, then we set the recurrence risk to be zero.

Note that if the family relationships are among only siblings, then the sibling recurrence risk estimate by QEM is exactly the same as Equation 9. Because sibling recurrence risk has been formally defined in two equivalent ways[35]:

- The probability that a sibling of an affected individual is also affected.

---

[1] $\hat{RR}_{2|1} = \begin{cases} \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_3+\hat{\beta}_5}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_3+\hat{\beta}_5}} & \text{if Father-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_4+\hat{\beta}_6}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_4+\hat{\beta}_6}} & \text{if Mother-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}}}{1+e^{\hat{\delta}+\hat{\gamma}}} & \text{if Child-Child.} \end{cases}$

[2] $\hat{RR}_{2|1} = \begin{cases} \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_1+\hat{\beta}_5}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_1+\hat{\beta}_5}} & \text{if Father-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_2+\hat{\beta}_6}}{1+e^{\hat{\delta}+\hat{\gamma}+\hat{\beta}_2+\hat{\beta}_6}} & \text{if Mother-Child} \\ \frac{e^{\hat{\delta}+\hat{\gamma}}}{1+e^{\hat{\delta}+\hat{\gamma}}} & \text{if Child-Child.} \end{cases}$

[3] $\hat{RR}_{race} = \begin{cases} \frac{e^{\hat{\delta}_W+\hat{\gamma}_W}}{1+e^{\hat{\delta}_W+\hat{\gamma}_W}} & \text{if whites} \\ \frac{e^{\hat{\delta}_O+\hat{\gamma}_O}}{1+e^{\hat{\delta}_O+\hat{\gamma}_O}} & \text{if other races.} \end{cases}$

- The proportion of affected individuals among all siblings of affected individuals in a population.

Since QEM completely specifies all possible probability combinations (i.e., $P(y_1 = 1, y_2 = 1)$, $P(y_1 = 1, y_2 = 0)$, $P(y_1 = 0, y_2 = 1)$, $P(y_1 = 0, y_2 = 0)$), $\hat{SRR}_{QEM} = P(y_1 = 1|y_2 = 1) = \frac{e^{\hat{\delta}+\hat{\gamma}}}{1+e^{\hat{\delta}+\hat{\gamma}}}$ can be derived based on the first definition under QEM. $\hat{SRR}_G$ is derived based on the second population-based definition which is a conditional proportion. Thus, $\hat{SRR}_{QEM} = \hat{SRR}_G$.

3.2.2 Prevalence and recurrence risk ratio calculations with complex survey sampling

Recurrence risk ratio is a statistic that normalizes recurrence risk dividing the recurrence risk by prevalence of disease. For recurrence risk ratio, we need to obtain an estimate of the prevalence of the disease from the data or use a value for the prevalence of disease from an outside source. Let *pr.pair* be the prevalence of the disease based on a network estimator for families with sibsizes larger than 1 as sibsizes equal to 1 do not contribute to recurrence risk; *pr.proband* means the prevalence of disease among the probands (y_{pi} denotes the disease status for the i^{th} sampled proband); and *pr.general* means the prevalence of disease based on network estimation from Graubard and Sirken (2013)[15], including all the sibling and proband information whether or not the sibsize is equal to 1.

To illustrate the calculation of the *pr.pair*, we use the original QEM. The different prevalence estimators are listed as follows,

$$\hat{pr}.pair = P(y_1 = 1) = P(y_1 = 1, y_2 = 0) + P(y_1 = y_2 = 1) = \frac{e^{\hat{\delta}} + e^{2\hat{\delta}+\hat{\gamma}}}{1 + 2e^{\hat{\delta}} + e^{2\hat{\delta}+\hat{\gamma}}},$$

$$\hat{pr}.proband = \frac{\sum_{i=1}^{n} w_i y_{pi}}{\sum_{i=1}^{n} w_i},$$

$$\hat{pr}.general = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} w_i}. \tag{10}$$

Since the reported disease information about siblings and probands can be different, it is possible to obtain different estimates between $\hat{pr}.proband$ and $\hat{pr}.general$. This can happen if there is differential reporting error in the reporting of disease by the proband for him/her self compared to report about disease stuatus of siblings. For example, the proband may not know the disease status of his/her sibling as well as their own disease status. Differences exist between $\hat{pr}.general$ and $\hat{pr}.pair$ if families with one child have a higher prevalence of the disease than families with more than one child. This can happen if the disease is strongly genetic and family sizes are limited when they have the gene. In addition, if sibsizes of all the families are larger than 1, then $\hat{pr}.general$ is the same as $\hat{pr}.pair$. We prefer using $pr.general$ as a prevalence estimator of disease, because it includes families with a single child and multiple children for the estimation based on siblings.

Therefore, the corresponding estimate of sibling recurrence risk ratio equals to

$$\hat{SRRR_G} = \frac{\hat{SRR}}{\hat{pr}.general}.$$

If $\hat{pr}.pair$ is used as the prevalence of disease, then

$$\hat{SRRR}_{QEM} = \frac{e^{\hat{\delta}+\hat{\gamma}}/(1+e^{\hat{\delta}+\hat{\gamma}})}{(e^{\hat{\delta}}+e^{2\hat{\delta}+\hat{\gamma}})/(1+2e^{\hat{\delta}}+e^{2\hat{\delta}+\hat{\gamma}})} = \frac{e^{\hat{\delta}}(1+2e^{\hat{\delta}}+e^{2\hat{\delta}+\hat{\gamma}})}{(1+e^{\hat{\delta}+\hat{\gamma}})^2}.$$

$\hat{SRRR}_{QEM}$ and $\hat{SRRR}_G$ are usually different since $\hat{pr}.general$ differs from $\hat{pr}.pair$ as described above. $\hat{SRRR}_G$ is preferred.

3.3 Variance estimator with complex survey data

As mentioned before, the estimator of middle part $Cov(\sum_i U_i(\Phi))$ in Equation 3[4] depends on the complex sampling designs. The variance estimators of recurrence risk estimates and recurrence risk ratio estimates are derived after obtaining the variance estimator of parameter estimates.

3.3.1 Variance estimator of parameter estimates

First, we show variance estimators for the parameter estimators in the QEM under SRS without replacement and with replacement respectively:

(1) SRS without replacement:

$$\hat{cov}(\sum_{i=1}^{n} U_i(\Phi)) = \frac{n}{N}\frac{N-n}{n-1}\sum_{i=1}^{n}(U_i(\hat{\Phi})-\bar{U}(\hat{\Phi}))(U_i(\hat{\Phi})-\bar{U}(\hat{\Phi}))^T.$$

[4] $\hat{cov}(\hat{\Phi}) = (\sum_{i=1}^{n}\frac{\partial U_i(\Phi)}{\partial\Phi}|_{\Phi=\hat{\Phi}})^{-1}\hat{cov}(\sum_{i=1}^{n}U_i(\Phi))((\sum_{i=1}^{n}\frac{\partial U_i(\Phi)}{\partial\Phi}|_{\Phi=\hat{\Phi}})^{-1})^T$.

(2) SRS with replacement:

$$\hat{cov}(\sum_{i=1}^{n} U_i(\Phi)) = \frac{n}{n-1}\sum_{i=1}^{n}(U_i(\hat{\Phi}) - \bar{U}(\hat{\Phi}))(U_i(\hat{\Phi}) - \bar{U}(\hat{\Phi}))^T.$$

Here, N is the finite population size and n is the sample size. $\bar{U}(\hat{\Phi})$ is the mean of score functions over n samples evaluated at $\Phi = \hat{\Phi}$, where $\Phi = (\delta, \gamma)^T$ and $\hat{\Phi} = (\hat{\delta}, \hat{\gamma})^T$ in the QEM.

The two major aspects of complex sampling designs is the use of stratification and clustering in sampling. The basic formulas for variance estimation are given by[22],

(3) Stratified SRS without replacement:

$$\hat{cov}(\sum_{h=1}^{H}\sum_{i=1}^{n_h} U_{hi}(\Phi)) = \sum_{h=1}^{H}\frac{n_h}{N_h}\frac{N_h - n_h}{n_h - 1}\sum_{i=1}^{n_h}(U_{hi}(\hat{\Phi}) - \bar{U_h}(\hat{\Phi}))(U_{hi}(\hat{\Phi}) - \bar{U_h}(\hat{\Phi}))^T.$$

(4) Stratified SRS with replacement:

$$\hat{cov}(\sum_{h=1}^{H}\sum_{i=1}^{n_h} U_{hi}(\Phi)) = \sum_{h=1}^{H}\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}(U_{hi}(\hat{\Phi}) - \bar{U_h}(\hat{\Phi}))(U_{hi}(\hat{\Phi}) - \bar{U_h}(\hat{\Phi}))^T,$$

where N_h and n_h represent the population size and sample size for h^{th} stratum. $U_{ij}(\hat{\Phi})$ is the j^{th} score function in the i^{th} sampled PSU, given $\Phi = \hat{\Phi}$.

(5) SRS of Equal-Sized PSUs with replacement:

$$\hat{cov}(\sum_{i=1}^{k}\sum_{j=1}^{n_0} U_{ij}(\Phi)) = \frac{n_0^2 k}{k-1}\sum_{i=1}^{k}(\bar{U}_i(\hat{\Phi}) - \bar{U}(\hat{\Phi}))(\bar{U}_i(\hat{\Phi}) - \bar{U}(\hat{\Phi}))^T.$$

Here, $\bar{U}_i(\hat{\Phi})$ is the mean of n_0 sampled individuals in the i^{th} sampled PSU evaluated at $\Phi = \hat{\Phi}$, and $\bar{U}(\hat{\Phi})$ is the overall mean of kn_0 sampled individuals when $\Phi = \hat{\Phi}$.

(6) SRS with replacement of PSUs of Differing Sizes: Suppose the finite population has K PSUs of differing sizes, N_1, N_2,..., N_K and a simple random sample of k PSUs is selected. And then a simple random sample of $n_i = \rho N_i$ individuals are sampled from the i^{th} selected PSU. Hence,

$$\hat{cov}(\sum_{i=1}^{k}\sum_{j=1}^{n_i} U_{ij}(\Phi)) = \frac{k}{k-1}\sum_{i=1}^{k} n_i^2(\bar{U}_i(\hat{\Phi}) - \bar{U}(\hat{\Phi}))(\bar{U}_i(\hat{\Phi}) - \bar{U}(\hat{\Phi}))^T.$$

(7) Stratified Multistage Sampling: Suppose that there are L strata with K_h PSUs in the population in the h^{th} stratum. From the h^{th} stratum, k_h PSUs are sampled with replacement with inclusion probabilities π_{h1}, π_{h2},...,π_{hk_h}. Then, SRS without replacement of n_{hi} individuals are sampled from N_{hi} individuals from the i^{th} sampled PSU of stratum h, and the sapmle weights for the n_{hi} sampled individuals are $w_{hij} = \frac{N_{hi}}{n_{hi}\pi_{hi}}$. Hence,

$$\hat{cov}(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}U_{hij}(\Phi)) = \sum_{h=1}^{L}\frac{k_h}{k_h-1}\sum_{i=1}^{k_h}(\bar{Y}_{hi}(\hat{\Phi}) - \bar{Y}(\hat{\Phi}))(\bar{Y}_{hi}(\hat{\Phi}) - \bar{Y}(\hat{\Phi}))^T,$$

where $\bar{Y}_{hi} = W_{hi}(\bar{U}_{hi} - \bar{U})$, $\bar{Y} = \frac{1}{k_h}\sum_{s=1}^{k_h} W_{hs}(\bar{U}_{hs} - \bar{U})$, and $W_{hi} = \sum_{j=1}^{n_{hi}} w_{hij}$ is the sum of sample weights of these sampled individuals. Here $\bar{U}_{hi} = \frac{\sum_{j=1}^{n_{hi}} w_{hij}U_{hij}}{\sum_{j=1}^{n_{hi}} w_{hij}}$ is the weighted mean of scores in the i^{th} sampled PSU of the h^{th} stratum, and $\bar{U} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}U_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}}$ is the overall weighted mean of scores[22].

Note in (6) and (7), we let the first-stage sampling of PSUs be with replacement sampling. Even if the actual sample selection in first stage is without replacement, the

first stage sampling is approximated in the variance estimation by with replacement sampling. This approximation is reasonable in most survey because the sampling fraction in the first stage is small.

As pointed out earlier, the NHIS is a stratified multistage cluster sample. Therefore, our variance estimation follows section (7) above; except the second stage involves more levels of sampling. We use the R package "survey" to calculate the variance esimators under different complex sampling methods.

3.3.2 Variance estimator of recurrence risk estimate

We make a parameter transformation to compute the variance estimator of recurrence risk estimate after obtaining the variance estimates of the parameter estimates. For instance, let $RR = \frac{e^{\hat{\delta}+\hat{\gamma}}}{1+e^{\hat{\delta}+\hat{\gamma}}}$, $T = e^{\hat{\delta}}$. Then, $\hat{\delta}$ and $\hat{\gamma}$ can be expressed as functions of (RR, T).

In addition, the method based on the paper by Graubard and Sirken(2013)[15] is another way to calculate the variance estimates of the estimated recurrence risk. Here is the formula shown in [15]:

$$\hat{var}(S\hat{R}R_G) \approx \hat{var}(\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}} z_{hij}) = \sum_{h=1}^{L}\sum_{i=i}^{t_h}(z_{hi} - \bar{z}_h)^2,$$

where $z_{hij} = w_{hij} \times (\hat{RR}) \times \left(\begin{array}{c} \frac{\frac{a_{hij}(a_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}}\frac{w_{hij}}{s_{hij}}a_{hij}(a_{hij}-1)} - \\ \frac{\frac{a_{hij}(s_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}}\frac{w_{hij}}{s_{hij}}a_{hij}(s_{hij}-1)} \end{array} \right)$, $z_{hi} = \sum_{j=1}^{t_{hi}} z_{hij}$, and $\bar{z}_h = \frac{1}{t_h}\sum_{i=1}^{t_h} z_{hi}$. Note that z_{hij} is the value for the j^{th} sampled individual in the i^{th}

PSU in the h^{th} stratum.

In Section 3.2.1, we have shown that $S\hat{R}R_{QEM}$ is equal to $S\hat{R}R_G$. Therefore, we would expect the the variance estimates will be the same too, i.e., $\hat{var}(S\hat{R}R_G) = \hat{var}(S\hat{R}R_{QEM})$. In Chapter 5, both of these methods will be used to calculate the recurrence risk estimates and variances for 1976 NHIS diabetes data.

3.3.3 Variance estimator of recurrence risk ratio estimate

The formula for the variance estimator of recurrence risk ratio estimate in Graubard and Sirken (2013) is:

$$\hat{var}(S\hat{RR}_G) \approx \hat{var}(\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}} x_{hij}) = \sum_{h=1}^{L}\sum_{i=i}^{t_h}(x_{hi} - \bar{x}_h)^2, \tag{11}$$

where $x_{hij} = w_{hij} \times (\hat{RR}) \times \left(\begin{array}{c} \frac{\frac{a_{hij}(a_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}}\frac{w_{hij}}{s_{hij}}a_{hij}(a_{hij}-1)} + \frac{1}{\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij}} - \\ \frac{\frac{a_{hij}(s_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}}\frac{w_{hij}}{s_{hij}}a_{hij}(s_{hij}-1)} - \frac{\frac{a_{hij}}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=i}^{t_h}\sum_{j=1}^{t_{hi}}\frac{w_{hij}}{s_{hij}}a_{hij}}) \end{array} \right)$,
$x_{hi} = \sum_{j=1}^{t_{hi}} x_{hij}$, and $\bar{x}_h = \frac{1}{t_h}\sum_{i=1}^{t_h} x_{hi}$ (Graubard and Sirken, 2013).

Since $\hat{pr}.general$[5] is preferred for calculation of the recurrence risk ratio, then Equation 11 will be used as the variance estimator of sibling recurrence risk ratio estimate. If all the families in the sample have sibsizes larger than 1, then $\hat{pr}.pair = \hat{pr}.general$ and hence $S\hat{RR}_G$[6] will be the same as $S\hat{RR}_{QEM}$[7]. If each estimator is the same, then we expect that the variance estimator of estimates, $\hat{var}(S\hat{RR}_G)$ and $\hat{var}(S\hat{RR}_{QEM})$, will be the same. Note that $S\hat{RR}_{QEM} = \frac{e^{\delta}(1+2e^{\hat{\delta}}+e^{2\hat{\delta}+\hat{\gamma}})}{(1+e^{\hat{\delta}+\hat{\gamma}})^2}$ is a

[5] $\hat{pr}.general = \frac{\sum_{i=1}^{n}\frac{w_i}{s_i}a_i}{\sum_{i=1}^{n}\frac{w_i}{s_i}s_i} = \frac{\sum_{i=1}^{n}\frac{w_i}{s_i}a_i}{\sum_{i=1}^{n}w_i}$.

[6] $S\hat{RR}_G = \frac{S\hat{RR}}{\hat{pr}.general}$.

[7] $S\hat{RR}_{QEM} = \frac{e^{\hat{\delta}}(1+2e^{\hat{\delta}}+e^{2\hat{\delta}+\hat{\gamma}})}{(1+e^{\hat{\delta}+\hat{\gamma}})^2}$.

function of $(\hat{\delta}, \hat{\gamma})$ and the variance estimator of $(\hat{\delta}, \hat{\gamma})$ has been obtained, therefore, $\hat{var}(\hat{SRRR}_{QEM})$ can be derived after parameter transformations.

3.4 Propensity score analysis

3.4.1 Background

Surveys are always observational studies. Therefore, when comparing recurrence risks or recurrence risk ratios between groups, e.g., racial groups, there can be important covariates (confounders) whose distribution differs between groups and are associated with the disease, e.g., age. Without controlling for a confounder, spurious group difference can result. We consider using propensity score methods for controlling for confounding.

Propensity score methods were proposed by Rosenbaum and Rubin in 1983[40], the propensity score is defined as the predicted probability, usually obtained from logistic regression of treatment group membership versus nontreatment group conditional on the observed covariates. These observed covariates usually are confounders for the analysis. Several researchers have suggested ways for using propensity scores in complex survey samples. Zanutto (Zanutto, 2006)[46] suggested that sample weights should be incorporated when estimating the treatment effects, but not included when estimating the propensity scores. Then in 2014, DuGoff et al. (DuGoff, Schuler and Stuart, 2014)[11] agreed with Zanutto but recommended that sample weights should be accounted for by including them as a covariate in the propensity score model. However, Ridgeway et al. (Ridgeway, Kovalchik and Griffin, 2015)[38] disagreed with Zanutto and DuGoff and concluded that the sample weights should be incorporated when fitting the logistic regression model as weights in the estimation of the regression coefficients and that the resulting propensity score weights should be combined

with the sample weights in the analysis.

After adjustment by propensity score weighting, the observed covariates between treatment group and nontreatment group should be similar. Propensity scores are usually obtained by using logistic regression.

We consider four different propensity score methods for balancing covariate distribution differences between groups: propensity score matching, stratification on the propensity score, covariate adjustment using the propensity score, and inverse propensity weighting [1]. The following is a brief introduction for these four methods.

3.4.1.1 Propensity score matching In randomized experiments, randomization assures in expectation unbiased estimation of treatment effects. However, for observational studies, there is no randomization. The purpose of propensity score matching is to create approximate randomization by selecting a sample of subjects that received a treatment that is comparable to sample or all of the subjects that are in the nontreatment group. The most common implementation of propensity score matching is one-to-one or pair matching, in which pairs of treated and nontreated subjects are formed, such that matched subjects have similar values of the propensity score. For the following two reasons, propensity score matching is usually used: (a) when very few units in the treatment group when comparing to the nontreated subjects (or visa versa); and (b) when selecting a subset of comparison units similar to the treatment unit is difficult because there are a large number of covariates. Nearest neighbor matching and nearest neighbor matching within a specified caliper distance of the propensity score (Rosenbaum and Rubin, 1985) are the most common methods for conducting the matching.

3.4.1.2 Stratification on the propensity score Stratification on the propensity score involves stratifying subjects into mutually exclusive subsets based on their different estimated propensity scores and then doing a stratified analysis. The bias will be reduced as the number of strata used increases, although the marginal reduction in bias decreases as the number of strata increases (Austin, 2011)[1]. Within each propensity score stratum, the subjects from treated and nontreated groups will have similar values of the propensity score. Therefore, when the propensity score has been correctly specified, the distribution of the observed covariates will be approximately similar between treated and nontreated groups within the same stratum.

3.4.1.3 Covariates adjustment Using this approach, the propensity score is included as a covariate in a regression analysis along with a treatment group indicator variable without the covariates used in estimating the propensity scores. As for the choice of regression model used for the analysis, for continuous outcomes, a linear model is usually chosen; for categorical outcomes, a logistic regression model could be chosen. The effect of treatment group is its corresponding regression coefficient in the fitted regression model.

3.4.1.4 Propensity score weighting This is the propensity score method we used for our research since our survey data and it is reasonable to obtain reccurence risk estimates by adjusting weights. Inverse probability of treatment weighting of the data using the propensity score creates weighted samples in which the distribution of measured observed covariate is independent of treatment variable. The use of propensity score weights is similar to the use of sample weights that are used to weighted survey

samples so that they are representative of the population[33]. There are two methods to compute the propensity score weights: average treatment effect (ATE)[19] and average treatment effect for the treated (ATT)[19]. Using either propensity score weighting, the product of sample weights and the propensity score weights is treated as new weights in weighted analysis. In the following example, we will compare the outcome analysis results before propensity score weighting and after propensity score weighting using logistic regression modeling to obtain the propensity scores to examine the propensity score adjustment.

As mentioned before, we use the propensity score weighting method to adjust for recurrence risk in our research. Thus, the new weight becomes the product of "sample weight" (sample weight of proband times network weight) and the propensity score weight[38]. The formulus are indicated as follows based on ATE and ATT respectively,

ATE:

$$\text{Propensity score weight} = \begin{cases} \frac{1}{\hat{f}(\text{trt variable=1}|\text{observed covariate})} & \text{if treatment group} \\ \frac{1}{\hat{f}(\text{trt variable=0}|\text{observed covariate})} & \text{if nontreatment group.} \end{cases}$$

ATT:

$$\text{Propensity score weight} = \begin{cases} 1 & \text{if treatment group} \\ \frac{\hat{f}(\text{treatment variable=1}|\text{observed covariate})}{\hat{f}(\text{treatment variable=0}|\text{observed covariate})} & \text{if nontreatment group.} \end{cases}$$

Note that $\hat{f}$ here is the predicted probability obtained from a logistic regression.

In the following example(Section 3.4.2), both ATE and ATT methods are used for propensity score weight calculation. Even though using propensity score weighting

is intended to remove imbalance in the covariates, distribution bias might still exist due to various reasons including lacking important covariates or misspecification of the propensity model. However, the propensity score weighting should reduce bias but it can increase variances of the treatment effect[1]. Additionally, even though we used propensity score weighting, applying other propensity score methods is an area of future research.

3.4.2 Simple example of propensity score weighting

In this section, a simple simulated example is performed to see how propensity score weighting can be performed. This example simulates an imbalance in the covariate age to compare whites and other races with regard to prevalence of a disease.

- Suppose treatment variable here is race ("trt"), and there are two races in the model: whites and other races. $trt = 1$ if race is whites; $trt = 0$ if race is others.
- The outcome is disease status ("Y"), where $Y \sim bernoulli(p)$, and $p = \frac{e^{\alpha+\lambda_1 \times age+\lambda_2 \times trt}}{1+e^{\alpha+\lambda_1 \times age+\lambda_2 \times trt}}$.
- Assume that $\alpha = -20$, $\lambda_1 = 0.5$, $\lambda_2 = 0.1$. Hence, outcome variable is related to both age and treatment variable race now.
- Use lognormal distribution to generate ages: $age \sim lognormal(log40, 0.1)$ if race is whites; $age \sim lognormal(log35, 0.1)$ if race is others.
- Population sizes are $N_W = 20000, N_O = 10000$ and sample sizes are $n_W = n_O = 600$ respectively, and the samples are SRS without replacement.

Hence, propensity score weighting calculations based on ATE and ATT methods are shown as follows using the simulated example data,

ATE:

$$\text{Propensity score weight} = \begin{cases} \frac{1}{\hat{f}(\text{trt=1}|\text{age})} & \text{if whites} \\ \frac{1}{\hat{f}(\text{trt=0}|\text{age})} & \text{if other races.} \end{cases}$$

ATT:

$$\text{Propensity score weight} = \begin{cases} 1 & \text{if whites} \\ \frac{\hat{f}(\text{trt=1}|\text{age})}{\hat{f}(\text{trt=0}|\text{age})} & \text{if other races.} \end{cases}$$

Note that new weight is the product of sample weight, that is $\frac{N_W}{n_W}$ for race whites and $\frac{N_O}{n_O}$ for other races, and propensity score weight. In subsequent parts, outcome analysis results and density plots before the propensity score weighting and after the propensity score weighting under ATE and ATT methods are shown respectively.

Figure 1: Example: Outcome Analysis before Propensity Score Analysis (ATE)

```
svyglm(formula = y ~ trt, design = des.str, family = quasibinomial("logit"))

Survey design:
svydesign(id = ~1, strata = ~group, data = data_s, weights = wt.str)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6582     0.1115  -14.88   <2e-16 ***
trt1          1.8186     0.1384   13.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: Example: Outcome Analysis after Propensity Score Analysis (ATE)

```
svyglm(formula = y ~ trt, design = des.str_ps, family = quasibinomial("logit"))

Survey design:
svydesign(id = ~1, strata = ~group, data = data_s, weights = wt.str_ps)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.28102    0.19564  -1.436    0.151
trt1        -0.03758    0.21570  -0.174    0.862

(Dispersion parameter for quasibinomial family taken to be 1.000834)
```

Figure 3: Example: Outcome Analysis before Propensity Score Analysis (ATT)

```
svyglm(formula = y ~ trt, design = des.str, family = quasibinomial("logit"))

Survey design:
svydesign(id = ~1, strata = ~group, data = data_s, weights = wt.str)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6582     0.1115  -14.88   <2e-16 ***
trt1          1.8186     0.1384   13.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Example: Outcome Analysis after Propensity Score Analysis (ATT)

```
svyglm(formula = y ~ trt, design = des.str_ps, family = quasibinomial("logit"))

Survey design:
svydesign(id = ~1, strata = ~group, data = data_s, weights = wt.str_ps)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.23056    0.22035   1.046    0.296
trt1        -0.07022    0.23511  -0.299    0.765

(Dispersion parameter for quasibinomial family taken to be 1.000834)
```

Figure 5: Example: Age Density before Propensity Score Analysis (ATE)

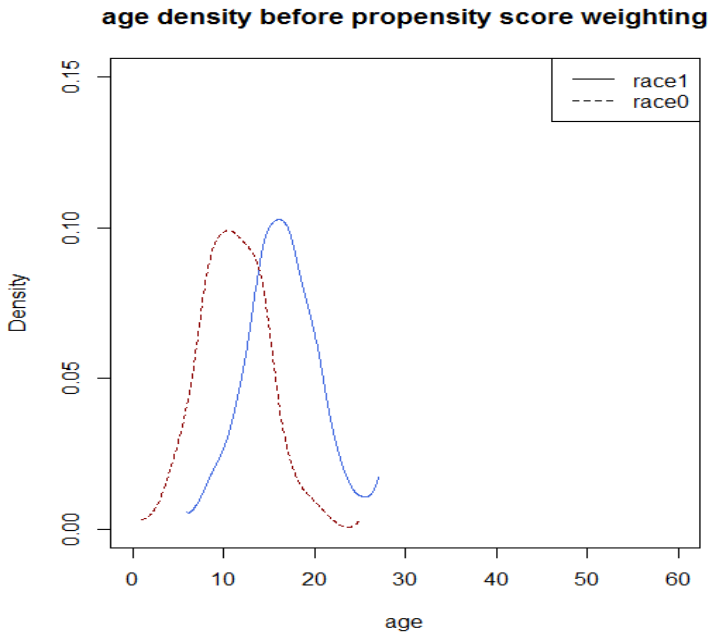


Figure 6: Example: Age Density after Propensity Score Analysis (ATE)

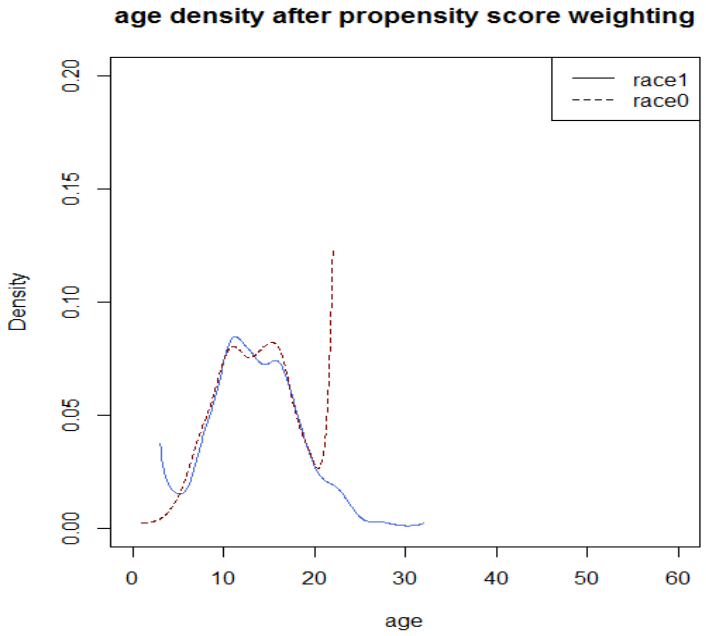


Figure 7: Example: Age Density before Propensity Score Analysis (ATT)

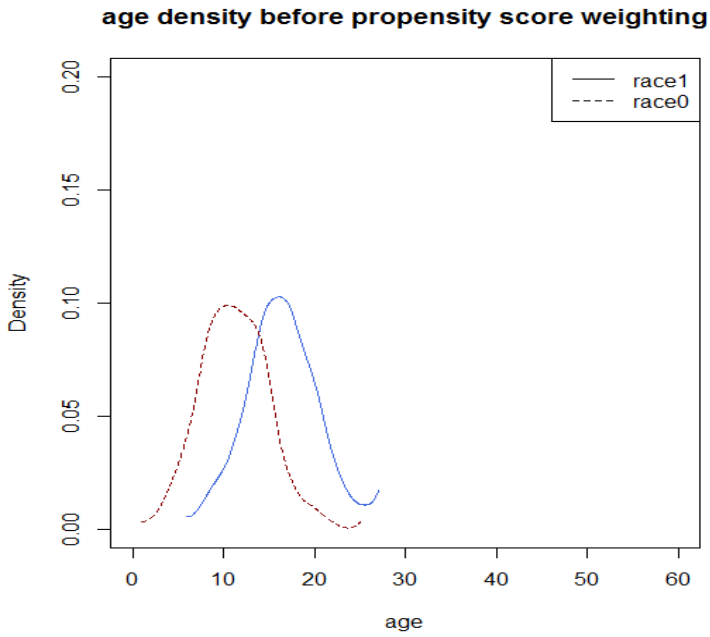
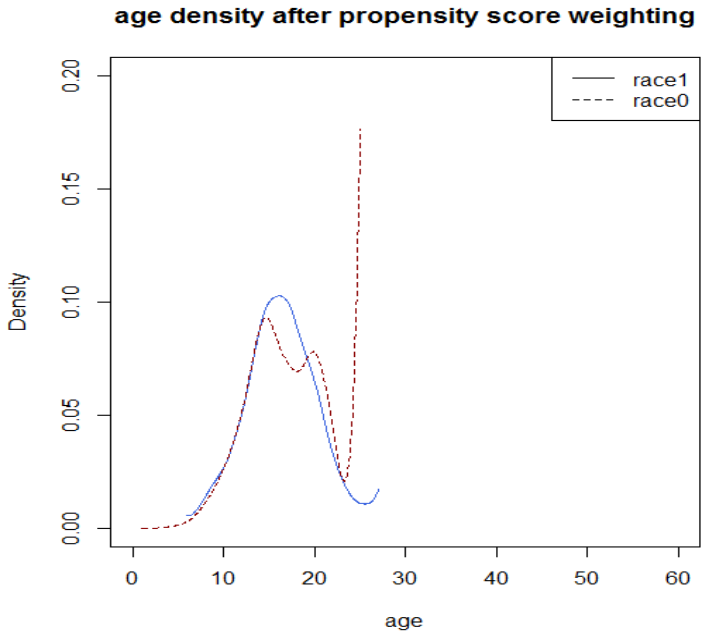


Figure 8: Example: Age Density after Propensity Score Analysis (ATT)



From Figure 1-Figure 4, we see that the treatment variable race is significantly associated with outcome Y before the propensity score weighting, however, it is not

significant with the outcome after propensity score weighting adjustment for age under both ATE and ATT methods.

From Figure 5-Figure 8 based on both ATE and ATT methods, the difference of age distributions between two races becomes smaller where the age densities overlap more after propensity score weighting adjustment than before the propensity score weighting. However, we found the distributions look abnormal after incorporating propensity score weights. This might be due to an edge effect. Hence, it is important to examine the distribution of propensity score weights. The followings are percentile summaries and boxplots based on propensity score weight only and the product of propensity score weight and sample weight, respectively. Note that only ATE method is used for the following examination.

Table 1: Example: Summary of propensity score weight distribution only (ATE)

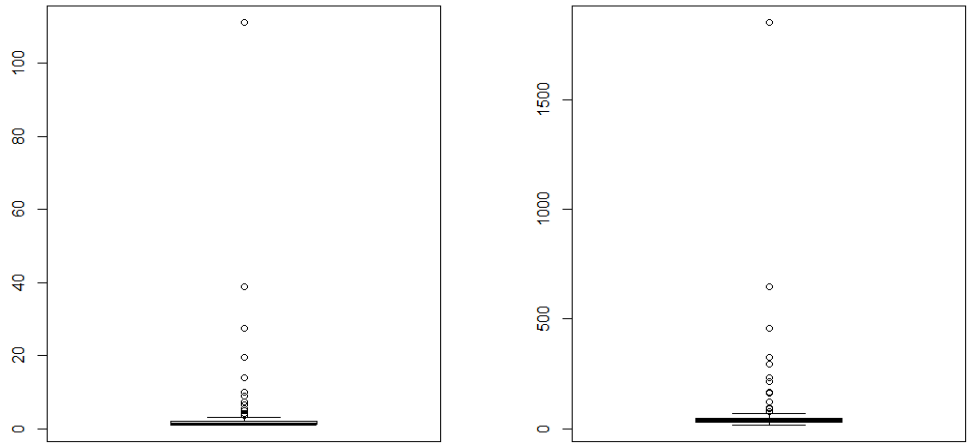
percentile	p.s. weight[8]	percentile	p.s. weight	percentile	p.s. weight
min	1.00	0.20	1.11	0.90	4.13
0.01	1.01	0.25	1.16	0.95	5.47
0.02	1.02	0.50	1.37	0.98	10.11
0.05	1.04	0.75	2.07	0.99	19.57
0.10	1.05	0.80	2.53	max	111.19

[8]p.s. weight is the propensity score weight.

Table 2: Example: Summary of weight product distribution (ATE)

percentile	weight product[9]	percentile	weight product	percentile	weight product
min	17.02	0.20	25.45	0.90	77.62
0.01	17.70	0.25	29.21	0.95	122.96
0.02	18.15	0.50	36.99	0.98	217.43
0.05	19.64	0.75	48.55	0.99	326.10
0.10	20.98	0.80	53.18	max	1853.14

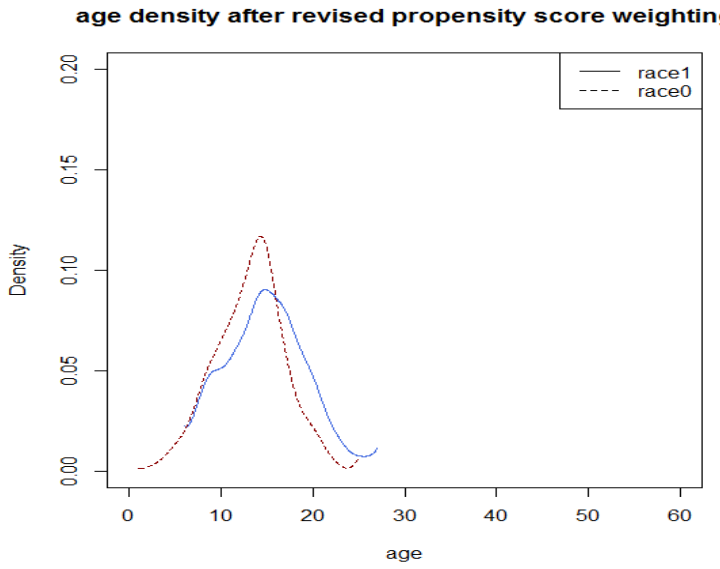
Figure 9: Example: Boxplot of propensity score weight (ATE)



Combining boxplot results and percentile analysis, we decided to trim the weights by using an upper value of weights, $Q_3 + 1.5IQR$, as a cutoff point and force the values of propensity score weight which are larger than this upper cutoff value to be equal to $Q_3 + 1.5IQR$. Then, age density is conducted again for different races using ATE method.

[9]weight product is the product of sample weight and propensity score weight.

Figure 10: Example: Revised Age Density after Propensity Score Analysis (ATE)



Now, the age density shows better overlap between race whites and other races without obvious edge effects. In conclusion, in this simple simulated example, propensity score weighting method balances the age distribution. The results are the same when using ATT method. In NHIS diabetes data, propensity score weighting will be applied to adjust age effect before studying the relationship between race and recurrence risk.

Chapter 4 Hypothesis testings with complex survey data

Wald test, score test and likelihood ratio test are well-known and important classical tests in statistical inference for simple random samples. Since likelihoods are difficult to determine in our setting with complex survey samples, classical score and likelihood ratio tests are not applicable. However, Wald and Quasi-score tests are available (Rao, 1998)[37] for complex survey data that are derived from weighted pseudo likelihoods. Both Wald and Quasi-score tests can take account of complex survey design features involving clustering and unequal sample weightings.

Wald tests have some limitations for simple random samples and complex samples. For instance, they are not invariant to nonlinear transformations of the parameters and often have poor behaviour if the sample size is small in terms of inflated Type I error[37]. In contrast, Quasi-score tests (including classical score tests) are invariant under reparametrization and can have better small sample properties than Wald tests[18] although they can be more difficult to derive.

In this chapter, we will explore both Wald tests and Quasi-score tests with complex survey data to test hypotheses about model parameters and examine Type I error under various simulated scenarios. The simulated examples are simple but aim to cover a range of hypothesis testing cases for parameters in our model. In next chapter, Wald tests and Quasi-score tests will be applied to 1976 NHIS diabetes data to test whether the association parameters or recurrence risks of diabetes are the same among different races or different family relationships.

4.1 Wald test

4.1.1 Theory

Wald test is named after a Hungarian statistician Abraham Wald, and it is a common method used to test the true values of parameters based on the sample estimates. Suppose $H_0 : \Phi_2 = \Phi_{20}$ using the sample data, where Φ is partitioned as $\Phi = (\Phi_1^T, \Phi_2^T)^T$ with Φ_2 a $s \times 1$ vector, and Φ_1 a $r \times 1$ vector. Then the test statistic is shown as follows,

$$X_W^2 = (\hat{\Phi}_2 - \Phi_{20})^T \hat{V}(\hat{\Phi}_2)^{-1}(\hat{\Phi}_2 - \Phi_{20}),$$

which is asymptotically a chi-square variable under null hypothesis with degrees of freedom s (Rao, Scott and Skinner, 1998). Here Φ denotes parameters, $\hat{\Phi}$ is parameter estimates and $\hat{V}(\Phi)$ is variance estimator of $\hat{\Phi}$. If it is a Wald test with linear constraints, e.g., $H_0 : H\Phi = q$ (H is a contrast matrix), then test statistic is given by,

$$X_W^2 = (H\hat{\Phi} - q)^T [H\hat{V}(\hat{\Phi})H^T]^{-1}(H\hat{\Phi} - q).$$

Note that X_W^2 under $H_0 : \Phi_2 = \Phi_{20}$ is a special case of X_W^2 under $H_0 : H\Phi = q$ (where $H = (0, 1)$ is partitioned in the same way as $\Phi = (\Phi_1^T, \Phi_2^T)^T$, and $q = \Phi_{20}$). We will use an F-distribution as a referent distribution for the Wald test, since it can perform better than a chi-square in approximating the asymptotic distribution of a Wald test statistic for data from complex sample surveys (Thomas and Rao, 1987)[44]. The F-distribution is given by,

$$F_W = \frac{(d-p+1)X_W^2}{d \times p},$$

and it follows an F-distribution under H_0 with degrees of freedom p and $d-p+1$ (Korn and Graubard, 1990), where p is the dimension of the parameters tested under the null hypothesis, and d is the number of PSUs in the sample minus the number of strata.

4.1.2 Type I error calculation

In this section, Type I error will be examined among different simulated scenarios. Here the simulated examples are based on simple random samples. Therefore, X_W^2 rather than F_W will be used here. In Chapter 6, we report simulation results with complex sampling involving stratified cluster sampling, so we will use F_W.

4.1.2.1 Ordinary-QEM (Scenario 1) Suppose there are only sibling relationships in the population. Thus, there are two parameters in the model: δ and γ, i.e., $\Phi = (\delta, \gamma)^T$. And we only focus on the families with size 2 and 3, even though our approach can be applied to any family sizes. Let the finite population sizes for families with size 2 and 3 be $N_2 = 20000$ and $N_3 = 10000$, respectively; while let the sample sizes be $n_2 = 600$ and $n_3 = 600$, respectively. The null hypothesis is $H_0 : \gamma = \gamma_0$, i.e., $\Phi_1 = \delta$ and $\Phi_2 = \gamma$. The purpose of this hypothesis test is to examine the value of parameter γ which represents the association between two siblings.

We have already shown how to calculate $\hat{\Phi}$ and $\hat{V}(\hat{\Phi})$ in Chapter 3. The following table displays Type I errors using simulated data under the null hypothesis.

Table 3: Example: Type I error (Wald Test) under $H_0 : \gamma = \gamma_0$

γ_0	replication times	Type I error
$\gamma_0 = 1.125$	1000	0.052
	5000	0.051
$\gamma_0 = -0.003$	1000	0.049
	5000	0.050

4.1.2.2 Relationship-EQEM There are eight parameters in relationship-EQEM: δ, γ, β_1, β_2, β_3, β_4, β_5 and β_6 (See Equation 4[10]), i.e., $\Phi = (\delta, \gamma, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T$. In the simulated example, let all the families have size 3, and finite population size $N_4 = 10000$ for families with relationships Father-Child-Child, $N_5 = 10000$ for families with relationships Mother-Child-Child, $N_6 = 30000$ for families with Child-Child-Child relationships and $N_7 = 20000$ for families with relationship Father-Mother-Child (same notation as Chapter 2), and we select $n_4 = n_5 = n_6 = n_7 = 1000$ samples from the families with four different relationships, respectively.

Part I (Scenario 2) Let $H_0 : \beta_2 = \beta_{20}$, i.e., $\Phi_1 = (\delta, \gamma, \beta_1, \beta_3, \beta_4, \beta_5, \beta_6)^T$ and $\Phi_2 = \beta_2$, then Type I error calculation is shown in following table:

[10] $P(y_1, y_2|I_1, I_2) = \frac{e^{(\delta+\beta_1 I_1+\beta_2 I_2)y_1+(\delta+\beta_3 I_1+\beta_4 I_2)y_2+(\gamma+\beta_5 I_1+\beta_6 I_2)y_1 y_2}}{1+e^{\delta+\beta_1 I_1+\beta_2 I_2}+e^{\delta+\beta_3 I_1+\beta_4 I_2}+e^{2\delta+\beta_1 I_1+\beta_2 I_2+\gamma+\beta_3 I_1+\beta_4 I_2+\beta_5 I_1+\beta_6 I_2}}$.

Table 4: Example: Type I error (Wald Test) under $H_0 : \beta_2 = \beta_{20}$

β_{20}	replication times	Type I error
$\beta_{20} = -0.017$	1000	0.043
	5000	0.051
$\beta_{20} = 0.437$	1000	0.041
	5000	0.047

Part II (Scenario 3) Let $H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix}$, i.e., $\Phi_1 = (\delta, \gamma, \beta_3, \beta_4, \beta_5, \beta_6)^T$ and $\Phi_2 = (\beta_1, \beta_2)^T$, then Type I error calculation is shown in following table:

Table 5: Example: Type I error (Wald Test) under $H_0 : \beta_1 = \beta_{10}$ and $H_0 : \beta_2 = \beta_{20}$

β_{10}, β_{20}	replication times	Type I error
$\beta_{10} = 0.438$, $\beta_{20} = 0.473$	1000	0.043
	5000	0.049
$\beta_{10} = 0.977$, $\beta_{20} = 0.891$	1000	0.041
	5000	0.047

4.1.2.3 Race-EQEM (Scenario 4) For race-EQEM, suppose there are two races in the model: whites and other races. Hence, there are four parameters: δ_W, γ_W, δ_O and γ_O (See Equation 7[11]), i.e., $\Phi = (\delta_W, \gamma_W, \delta_O, \gamma_O)^T$. Let the finite population size $N_{W2} = 20000$ for families with race white, $N_{O2} = 10000$ for families with other races, and the sample sizes be $n_{W2} = n_{O2} = 600$ (same notation as Chapter 2),

[11] $P(y_1, y_2|race = W) = \frac{e^{\delta_W(y_1+y_2)+\gamma_W y_1 y_2}}{1+2e^{\delta_W}+e^{2\delta_W+\gamma_W}}$; $P(y_1, y_2|race = O) = \frac{e^{\delta_O(y_1+y_2)+\gamma_O y_1 y_2}}{1+2e^{\delta_O}+e^{2\delta_O+\gamma_O}}$.

respectively. Hence, Type I error under $H_0 : \gamma_W = \gamma_O$ (Note that $\Phi_1 = (\delta_W, \delta_O)^T$ and $\Phi_2 = (\gamma_W, \gamma_O)^T$) is given by,

Table 6: Example: Type I error (Wald Test) under $H_0 : \gamma_W = \gamma_O$

γ_W, γ_O	replication times	Type I error
$\gamma_W = \gamma_O = 0.013$	1000	0.042
	5000	0.052
$\gamma_W = \gamma_O = 0.023$	1000	0.041
	5000	0.048

4.2 Quasi-Score test

4.2.1 Theory

For a finite population, suppose the score function is $S(\Phi) = \sum_{i=1}^{N} U_i(\Phi)$. We use $\hat{S}(\Phi) = \sum_{i=1}^{n} w_i U_i(\Phi)$ to estimate the score function for finite population, and set it to be zero to obtain the parameter estimate $\hat{\Phi}$. Here, $\Phi = (\Phi_1^T, \Phi_2^T)^T$ is the same as used for the Wald test.

Under null hypothesis $H_0 : \Phi_2 = \Phi_{20}$, the test statistic for Quasi-score test is given by,

$$X_S^2 = \widetilde{S}_2^T \widetilde{V}_{2S}^{-1} \widetilde{S}_2,$$

where $\widetilde{\Phi} = (\widetilde{\Phi}_1^T, \Phi_{20}^T)^T$ is the solution of $\hat{S}_1(\widetilde{\Phi}) = 0$, and $\hat{S} = (\hat{S}_1^T, \hat{S}_2^T)^T$ is partitioned in the same way as Φ[37].

There is another test statistic formula of Quasi-score test, e.g., $H_0 : \gamma_W = \gamma_O$. The test statistic is given by (Boos, 1992),

$$X_S^2 = \hat{S}(\widetilde{\Phi})^T \widetilde{I}_S^{-1} \widetilde{H}^T (\widetilde{H} \widetilde{I}_S^{-1} \widetilde{V}_S \widetilde{I}_S^{-1} \widetilde{H}^T)^{-1} \widetilde{H} \widetilde{I}_S^{-1} \hat{S}(\widetilde{\Phi}),$$

where $\widetilde{\Phi}$ minimizes log-pseudo weighted likelihood subject to the constraints[6]

$$S(\widetilde{\Phi}) - H(\widetilde{\Phi})^T \widetilde{\lambda} = 0, \ h(\widetilde{\Phi}) = 0,$$

where $\widetilde{I}_S = -\frac{\partial S(\Phi)}{\partial \Phi}$, evaluated at $\Phi = \widetilde{\Phi}$, and $\widetilde{V}_S$ is a consistent variance estimator of $\hat{S}(\widetilde{\Phi})$ which we will use a jackknife variance estimator. Here, $h(\Phi) = \gamma_W - \gamma_O$, $H(\Phi) = \frac{\partial h(\Phi)}{\partial \Phi} = (0, 0, 1, -1)$, a 1×4 vector, and λ is the Lagrange multiplier.

Similar to the Wald test, X_S^2 under $H_0 : \Phi_2 = \Phi_{20}$ is a special case of X_S^2 under $H_0 : H\Phi = q$. We will use an F-distribution, a reference distribution to the Quasi-score test for complex survey sample, which is given by,

$$F_S = \frac{(d - p + 1) X_S^2}{d \times p},$$

and it follows an F distribution under H_0 with degrees of freedom p and $d - p + 1$ (Rao, Scott and Skinner, 1998).

The jackknife has the advantage over Taylor linearization variance estimation that it can easily incorporate post-stratification and unit nonresponse adjustments of the sample weights in the estimation for $\widehat{V}_S$[22].

4.2.2 Type I error calculation

We use the same set up as the Wald test for each scenario.

4.2.2.1 Ordinary-QEM (Scenario 1) Under $H_0 : \gamma = \gamma_0$,

$$\begin{aligned}
\hat{S}_1(\widetilde{\Phi}) &= \sum_{i=1}^{n_2} \frac{N_2}{n_2}(y_{i1} + y_{i2}) - N_2 \frac{2e^{\delta} + 2e^{2\delta+\gamma}}{1 + 2e^{\delta} + e^{2\delta+\gamma}} \\
&\quad + 2\sum_{i=1}^{n_3} \frac{N_3}{n_3}(y_{i1} + y_{i2} + y_{i3}) - 3N_3 \frac{2e^{\delta} + 2e^{2\delta+\gamma}}{1 + 2e^{\delta} + e^{2\delta+\gamma}}, \\
\hat{S}_2(\widetilde{\Phi}) &= \sum_{i=1}^{n_2} \frac{N_2}{n_2} y_{i1}y_{i2} - N_2 \frac{e^{2\delta+\gamma}}{1 + 2e^{\delta} + e^{2\delta+\gamma}} \\
&\quad + \sum_{i=1}^{n_3} \frac{N_3}{n_3}(y_{i1}y_{i2} + y_{i1}y_{i3} + y_{i2}y_{i3}) - 3N_3 \frac{e^{2\delta+\gamma}}{1 + 2e^{\delta} + e^{2\delta+\gamma}}.
\end{aligned}$$

.

We set $\hat{S}_1(\widetilde{\Phi}) = 0$ to obtain $\widetilde{\Phi} = (\widetilde{\delta}, \gamma_0)^T$, and then plug $\widetilde{\Phi}$ into $\hat{S}_2(\widetilde{\Phi})$ to get $\widetilde{S}_2$.

Hence, Type I error is displayed in following table:

Table 7: Example: Type I error (Quasi-Score Test) under $H_0 : \gamma = \gamma_0$

γ_0	replication times	Type I error
$\gamma_0 = 0.014$	1000	0.046
	5000	0.051
$\gamma_0 = 1.125$	1000	0.046
	5000	0.048

4.2.2.2 Relationship-EQEM Based on results from Chapter 2 and Chapter 3, the score functions under relationship-EQEM are given as follows,

$$\begin{aligned}
S(\delta) = & \frac{2N_4}{n_4}\sum_{i=1}^{n_4}(y_{i1}+y_{i2}+y_{i3}) + \frac{2N_5}{n_5}\sum_{i=1}^{n_5}(y_{i1}+y_{i2}+y_{i3}) \\
& + \frac{2N_6}{n_6}\sum_{i=1}^{n_6}(y_{i1}+y_{i2}+y_{i3}) + \frac{N_7}{n_7}\sum_{i=1}^{n_7}(y_{i1}+y_{i3}+y_{i2}+y_{i3}) \\
& - (2N_4+N_7)\frac{e^{\delta+\beta_1}+e^{\delta+\beta_3}+2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}} \\
& - (2N_5+N_7)\frac{e^{\delta+\beta_2}+e^{\delta+\beta_4}+2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1+e^{\delta+\beta_2}+e^{\delta+\beta_4}+e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}} \\
& - (N_4+N_5+3N_6)\frac{2e^{\delta}+2e^{2\delta+\gamma}}{1+2e^{\delta}+e^{2\delta+\gamma}},
\end{aligned}$$

$$\begin{aligned}
S(\gamma) = & \frac{N_4}{n_4}\sum_{i=1}^{n_4}(y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3}) + \frac{N_5}{n_5}\sum_{i=1}^{n_5}(y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3}) \\
& + \frac{N_6}{n_6}\sum_{i=1}^{n_6}(y_{i1}y_{i2}+y_{i1}y_{i3}+y_{i2}y_{i3}) + \frac{N_7}{n_7}\sum_{i=1}^{n_7}(y_{i1}y_{i3}+y_{i2}y_{i3}) \\
& - (2N_4+N_7)\frac{e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}} \\
& - (2N_5+N_7)\frac{e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1+e^{\delta+\beta_2}+e^{\delta+\beta_4}+e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}} \\
& - (N_4+N_5+3N_6)\frac{e^{2\delta+\gamma}}{1+2e^{\delta}+e^{2\delta+\gamma}},
\end{aligned}$$

$$\begin{aligned}
S(\beta_1) = & \frac{2N_4}{n_4}\sum_{i=1}^{n_4}y_{i1} + \frac{N_7}{n_7}\sum_{i=1}^{n_7}y_{i1} \\
& - (2N_4+N_7)\frac{e^{\delta+\beta_1}+2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1+e^{\delta+\beta_1}+e^{\delta+\beta_3}+e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},
\end{aligned}$$

$$S(\beta_2) = \frac{2N_5}{n_5}\sum_{i=1}^{n_5} y_{i1} + \frac{N_7}{n_7}\sum_{i=1}^{n_7} y_{i2} - (2N_5 + N_7)\frac{e^{\delta+\beta_2} + 2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}},$$

$$S(\beta_3) = \frac{2N_4}{n_4}\sum_{i=1}^{n_4} y_{i2} + \frac{N_7}{n_7}\sum_{i=1}^{n_7} y_{i3} - (2N_4 + N_7)\frac{e^{\delta+\beta_3} + 2e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},$$

$$S(\beta_4) = \frac{2N_5}{n_5}\sum_{i=1}^{n_5} y_{i2} + \frac{N_7}{n_7}\sum_{i=1}^{n_7} y_{i3} - (2N_5 + N_7)\frac{e^{\delta+\beta_4} + 2e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}},$$

$$S(\beta_5) = \frac{N_4}{n_4}\sum_{i=1}^{n_4}(y_{i1}y_{i2} + y_{i1}y_{i3}) + \frac{N_7}{n_7}\sum_{i=1}^{n_7} y_{i1}y_{i3} - (2N_4 + N_7)\frac{e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}}{1 + e^{\delta+\beta_1} + e^{\delta+\beta_3} + e^{2\delta+\gamma+\beta_1+\beta_3+\beta_5}},$$

$$S(\beta_6) = \frac{N_5}{n_5}\sum_{i=1}^{n_5}(y_{i1}y_{i2} + y_{i1}y_{i3}) + \frac{N_7}{n_7}\sum_{i=1}^{n_7} y_{i2}y_{i3} - (2N_5 + N_7)\frac{e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}{1 + e^{\delta+\beta_2} + e^{\delta+\beta_4} + e^{2\delta+\gamma+\beta_2+\beta_4+\beta_6}}.$$

Part I (Scenario 2) Under $H_0 : \beta_2 = \beta_{20}$, $\Phi_1 = (\delta, \gamma, \beta_1, \beta_3, \beta_4, \beta_5, \beta_6)^T$ and $\Phi_2 = \beta_2$.

$$\hat{S}_1(\widetilde{\Phi}) = \begin{pmatrix} S(\delta) \\ S(\gamma) \\ S(\beta_1) \\ S(\beta_3) \\ S(\beta_4) \\ S(\beta_5) \\ S(\beta_6) \end{pmatrix}.$$

And,

$$\hat{S}_2(\widetilde{\Phi}) = S(\beta_2).$$

We set $\hat{S}_1(\widetilde{\Phi}) = 0$ to obtain $\widetilde{\Phi} = (\widetilde{\Phi_1}, 0)^T$, and then we plug $\widetilde{\Phi}$ into $\hat{S}_2(\widetilde{\Phi})$ to get $\widetilde{S}_2$ needed in the Quasi-score test statistic calculation.

We will calculate the Type I error using the simulated data under $H_0 : \beta_2 = \beta_{20}$. The results are given by,

Table 8: Example: Type I error (Quasi-Score Test) under $H_0 : \beta_2 = \beta_{20}$

β_{20} under null hypothesis	simulation times	Type I error
$\beta_{20} = 0.437$	1000	0.063
	5000	0.053
$\beta_{20} = 0.352$	1000	0.061
	5000	0.055

Part II (Scenario 3) Let $H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix}$, so $\Phi_1 = (\delta, \gamma, \beta_3, \beta_4, \beta_5, \beta_6)^T$ and $\Phi_2 = (\beta_1, \beta_2)^T$ in this case. Then,

$$\hat{S}_1(\widetilde{\Phi}) = \begin{pmatrix} S(\delta) \\ S(\gamma) \\ S(\beta_3) \\ S(\beta_4) \\ S(\beta_5) \\ S(\beta_6) \end{pmatrix}.$$

And,

$$\hat{S}_2(\widetilde{\Phi}) = \begin{pmatrix} S(\beta_1) \\ S(\beta_2) \end{pmatrix}.$$

We set $\hat{S}_1(\widetilde{\Phi}) = 0$ to obtain $\widetilde{\Phi} = (\widetilde{\Phi_1}, 0)^T$, and then we plug $\widetilde{\Phi}$ into $\hat{S}_2(\widetilde{\Phi})$ to obtain $\widetilde{S}_2$. And then, Type I error is calculated as follows:

Table 9: Example: Type I error (Quasi-Score Test) under $H_0 : \beta_1 = \beta_{10}$ and $H_0 : \beta_2 = \beta_{20}$

β_{10}, β_{20} under null hypothesis	simulation times	Type I error
$\beta_{10} = 1.215$, $\beta_{20} = 1.163$	1000	0.063
	5000	0.053
$\beta_{10} = 0.438$, $\beta_{20} = 0.473$	1000	0.061
	5000	0.054

4.2.2.3 Race-EQEM (Scenario 4) We use the same set up as the Wald test, and there are four parameters in the model: δ_W, γ_W, δ_O and γ_O. Under $H_0 : \gamma_W = \gamma_O$, we have

$$\hat{S}_1(\widetilde{\Phi}) = \begin{pmatrix} \hat{S}_1(\widetilde{\Phi})_{(1)} \\ \hat{S}_1(\widetilde{\Phi})_{(2)} \end{pmatrix}.$$

And,

$$\hat{S}_2(\widetilde{\Phi}) = \begin{pmatrix} \hat{S}_2(\widetilde{\Phi})_{(1)} \\ \hat{S}_2(\widetilde{\Phi})_{(2)} \end{pmatrix},$$

where

$$\hat{S}_1(\widetilde{\Phi})_{(1)} = 2\sum_{i=1}^{n_{W2}} \frac{N_{W2}}{n_{W2}}(y_{i1} + y_{i2} + y_{i3}) - 3N_{W2}\frac{2e^{\delta_W} + 2e^{2\delta_W+\gamma_W}}{1 + 2e^{\delta_W} + e^{2\delta_W+\gamma_W}}, \tag{12}$$

$$\hat{S}_1(\widetilde{\Phi})_{(2)} = 2\sum_{i=1}^{n_{O2}} \frac{N_{O2}}{n_{O2}}(y_{i1} + y_{i2} + y_{i3}) - 3N_{O2}\frac{2e^{\delta_O} + 2e^{2\delta_O+\gamma_O}}{1 + 2e^{\delta_O} + e^{2\delta_O+\gamma_O}}, \tag{13}$$

$$\hat{S}_2(\widetilde{\Phi})_{(1)} = \sum_{i=1}^{n_{W2}} \frac{N_{W2}}{n_{W2}}(y_{i1}y_{i2} + y_{i1}y_{i3} + y_{i2}y_{i3}) - 3N_{W2}\frac{e^{2\delta_W+\gamma_W}}{1 + 2e^{\delta_W} + e^{2\delta_W+\gamma_W}}, \tag{14}$$

$$\hat{S}_2(\widetilde{\Phi})_{(2)} = \sum_{i=1}^{n_{O2}} \frac{N_{O2}}{n_{O2}}(y_{i1}y_{i2} + y_{i1}y_{i3} + y_{i2}y_{i3}) - 3N_{O2}\frac{e^{2\delta_O+\gamma_O}}{1 + 2e^{\delta_O} + e^{2\delta_O+\gamma_O}}. \tag{15}$$

And

$$S(\widetilde{\Phi}) = \begin{pmatrix} \hat{S}_1(\widetilde{\Phi}) \\ \hat{S}_2(\widetilde{\Phi}) \end{pmatrix}$$

is a 4×1 vector.

Under constraint $\gamma_W = \gamma_O$, we have[6]

$$S(\widetilde{\Phi}) = \begin{pmatrix} \hat{S}_1(\widetilde{\Phi})_{(1)} \\ \hat{S}_1(\widetilde{\Phi})_{(2)} \\ \hat{S}_2(\widetilde{\Phi})_{(1)} \\ \hat{S}_2(\widetilde{\Phi})_{(2)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \lambda \\ -\lambda \end{pmatrix}. \tag{16}$$

Plugging Equations 12, 13, 14 and 15 into Equation 16, we can obtain that $\widetilde{\Phi} = (\widetilde{\delta_W}, \widetilde{\delta_O}, \widetilde{\gamma_W}, \widetilde{\gamma_O})^T$, note that $\widetilde{\gamma_W} = \widetilde{\gamma_O}$ in this scenario, and Lagrange multiplier λ will be deleted after summarizing $\hat{S}_2(\widetilde{\Phi})_{(1)}$ and $\hat{S}_2(\widetilde{\Phi})_{(2)}$ in Equation 16.

Next step is to figure out I_S, and the derivatives are calculated as follows,

$$I_S = \begin{pmatrix} -\frac{\partial \hat{S}_1}{\partial \delta_W} & -\frac{\partial \hat{S}_1}{\partial \delta_O} & -\frac{\partial \hat{S}_1}{\partial \gamma_W} & -\frac{\partial \hat{S}_1}{\partial \gamma_O} \\ -\frac{\partial \hat{S}_2}{\partial \delta_W} & -\frac{\partial \hat{S}_2}{\partial \delta_O} & -\frac{\partial \hat{S}_2}{\partial \gamma_W} & -\frac{\partial \hat{S}_2}{\partial \gamma_O} \end{pmatrix} = \begin{pmatrix} I_{(1)} & I_{(2)} & I_{(3)} & I_{(4)} \\ I_{(5)} & I_{(6)} & I_{(7)} & I_{(8)} \\ I_{(9)} & I_{(10)} & I_{(11)} & I_{(12)} \\ I_{(13)} & I_{(14)} & I_{(15)} & I_{(16)} \end{pmatrix},$$

where I_S is a 4×4 matrix, and

$$I_{(1)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \delta_W} = 6N_{W2}\frac{e^{\delta_W}(1 + e^{2\delta_W + \gamma_W} + 2e^{\delta_W + \gamma_W})}{(1 + 2e^{\delta_W} + e^{2\delta_W + \gamma_W})^2},$$

$$I_{(2)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \delta_O} = 0,$$

$$I_{(3)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \gamma_W} = 6N_{W2}\frac{e^{2\delta_W+\gamma_W}(1+e^{\delta_W})}{(1+2e^{\delta_W}+e^{2\delta_W+\gamma_W})^2},$$

$$I_{(4)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \gamma_O} = 0,$$

$$I_{(5)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(2)}}{\partial \delta_W} = 0,$$

$$I_{(6)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \delta_O} = 6N_{O2}\frac{e^{\delta_O}(1+e^{2\delta_O+\gamma_O}+2e^{\delta_O+\gamma_O})}{(1+2e^{\delta_O}+e^{2\delta_O+\gamma_O})^2},$$

$$I_{(7)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(2)}}{\partial \gamma_W} = 0,$$

$$I_{(8)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \gamma_O} = 6N_{O2}\frac{e^{2\delta_O+\gamma_O}(1+e^{\delta_O})}{(1+2e^{\delta_O}+e^{2\delta_O+\gamma_O})^2},$$

$$I_{(9)} = -\frac{\partial \hat{S}_2(\widetilde{\Phi})_{(1)}}{\partial \delta_W} = 6N_{W2}\frac{e^{2\delta_W+\gamma_W}(1+e^{\delta_W})}{(1+2e^{\delta_W}+e^{2\delta_W+\gamma_W})^2},$$

$$I_{(10)} = -\frac{\partial \hat{S}_2(\widetilde{\Phi})_{(1)}}{\partial \delta_O} = 0,$$

$$I_{(11)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \gamma_W} = 3N_{W2}\frac{e^{2\delta_W+\gamma_W}(1+2e^{\delta_W})}{(1+2e^{\delta_W}+e^{2\delta_W+\gamma_W})^2},$$

$$I_{(12)} = -\frac{\partial \hat{S}_2(\widetilde{\Phi})_{(1)}}{\partial \gamma_O} = 0,$$

$$I_{(13)} = -\frac{\partial \hat{S}_2(\widetilde{\Phi})_{(2)}}{\partial \delta_W} = 0,$$

$$I_{(14)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \delta_O} = 6N_{O2}\frac{e^{2\delta_O+\gamma_O}(1+e^{\delta_O})}{(1+2e^{\delta_O}+e^{2\delta_O+\gamma_O})^2},$$

$$I_{(15)} = -\frac{\partial \hat{S}_2(\widetilde{\Phi})_{(2)}}{\partial \gamma_W} = 0,$$

$$I_{(16)} = -\frac{\partial \hat{S}_1(\widetilde{\Phi})_{(1)}}{\partial \gamma_O} = 3N_{O2}\frac{e^{2\delta_O+\gamma_O}(1+2e^{\delta_O})}{(1+2e^{\delta_O}+e^{2\delta_O+\gamma_O})^2},$$

given that $\widetilde{\Phi} = (\widetilde{\delta_W}, \widetilde{\delta_O}, \widetilde{\gamma_W}, \widetilde{\gamma_O})^T$.

The Type I error is obtained as follows:

Table 10: Example: Type I error (Quasi-Score Test) under $H_0 : \gamma_W = \gamma_O$

γ_W, γ_B under null hypothesis	simulation times	Type I error
$\gamma_W = \gamma_B = 0.014$	1000	0.062
	5000	0.053
$\gamma_W = \gamma_B = -0.001$	1000	0.060
	5000	0.054

In summary, from all above scenarios $1-4$ based on both Wald test and Quasi-score test, we find out that the Type I errors are closer to 0.05 especially when the

number of replications increases. And usually Wald tests are more conservative than Quasi-score tests.

Also, Wald tests for testing recurrence risks of disease can be constructed since parameter estimates and variance estimators of recurrence risk of disease have been derived in Chapter 3.

Chapter 5 1976 NHIS diabetes data analysis

In this chapter, we will apply our methods to 1976 NHIS diabetes data, estimate and conduct inferences on familial recurrence risk of diabetes and its relationship with covariates, e.g., type of family relationship and race.

The NHIS has been conducted annually by the National Center for Health Statistics (NCHS), Center for Disease Control and Prevention (CDC) since 1957 to obtain health and health services data on the civilian noninstitutionalized population of the United States. Staff from U.S. Bureau of the Census interview the households face-to-face after randomly selecting these households.

The sample design is multistage. Firstly, the whole population of the United States was divided into geographically defined areas called primary sampling units (PSUs), which are counties, contiguous counties, small cities or parts of larger cities and collectively cover the 50 states and the District of Columbia. Next, the PSUs are grouped into strata. After randomly sampling PSUs, small, compact clusters of housing units are then selected within each sampled PSU, called segments. After that, several housing units are randomly sampled from each sampled segment. Finally, individuals are randomly sampled within each sampled housing unit, see https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm for further details about sample design of the NHIS.

The 1976 NHIS diabetes data incorporates the household structure, diabetes disease information and other characteristics of each individual from each sampled household. There are $113,178$ sampled individuals with observations and 180 variables in total. Because network sample estimation and at times propensity score weighting will be used, the weight used in the estimation is equal to the product of the sample

weight for proband, the network weight and the propensity score weight when the propensity weight is used.

In conducting the data analysis for 1976 NHIS diabetes data, we assume that the number of living siblings reported in survey live inside the United States, and are civilian noninstitutionalized. This assumption is needed in order to calculate the network sample weights.

5.1 Original-QEM

For the original-QEM, we only consider sibling relationships with no parameters for race, therefore, there are only two parameters δ and γ in our model. In the next section 5.1.1, we study data reported about living siblings only, and in section 5.1.2, we study data reprted about both alive and dead siblings, to estimate recurrence risk, recurrence risk ratio and calculate corresponding variance estimators.

5.1.1 Using data reported about only living siblings

There are four variables needed: proband diabetes status (Col 509 on the public use file[12]) which is a 0/1 binary variable; number of living siblings (Col 510-511); number of living siblings with diabetes (Col 512-513); sample weight for the proband (Col 188-192). The followings are the frequency tables for the number of living siblings and the living siblings with diabetes, respectively:

[12] https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.

Table 11: 1976 NHIS Data: Frequencies for the number of living siblings

living siblings, No.	freq	living siblings, No.	freq	living siblings, No.	freq
0	14115	9	1445	18	3
1	25502	10	896	19	2
2	22923	11	591	20	3
3	16660	12	278	24	1
4	11162	13	170	60	1
5	7420	14	66	62	1
6	4824	15	41	83	1
7	3329	16	9	unknown	1520[13]
8	2204	17	11		

Table 12: 1976 NHIS Data: Frequencies for the number of affected living siblings

affected(diabetic) living siblings, No.	freq	affected living siblings, No.	freq	affected living siblings, No.	freq
0	107234	4	40	9	1
1	3693	5	15	10	1
2	548	6	4	12	1
3	118	7	3	unknown	1520

Note that the frequencies in Table 11 and Table 12 indicate the unweighted count of probands, not families, since it is possible that two or more siblings in the same

[13]There are 1520 probands reporting "unknown" for the number of their living siblings. We treat them as missing values and they are not included in our study.

family were sampled in the survey. Note that four probands report having more than 20 siblings. In Table 11, there is a drop off in the number of living siblings from 15 to 16. Therefore, we assumed that reports of the number of living siblings that were greater than or equal to 16 were likely reporting errors, and we deleted these observations.

According to the usernotes of 1976 NHIS diabetes data, many of the original sample design variables are suppressed in the NHIS public use files due to confidentially issues, but NCHS released public use sample design variables representing pseudo-strata and pseudo-PSU variables to be used for variance estimation in analysis of the data. These are constructed variables for analytic purposes that do not reflect the actual stratification and primary sampling units used to draw the samples. Using these variables in the variance estimation performs reasonably well as an alternative to using the original design variables. In 1976 NHIS data, there are 2 pseudo-PSUs within each pseudo-stratum, and there are 149 pseudo-strata for a total of 298 pseudo-PSUs.

After data editing, we applied the original-QEM to the NHIS data, obtaining following results:

Table 13: 1976 NHIS Data: Estimation results for original-QEM(only living siblings)

	parameter estimate (var est)		RR & RRR estimate [14] (var est)
δ	-4.13 (0.00026)	RR(original-QEM)	0.1545 (0.0000447)
γ	2.43 (0.00290)	RRR	7.74 (0.0994)

The RR estimate and variance in Table 13 agree with the estimated results in Graubard and Sirken (2013)[15]. Note that for estimating recurrence risk ratio, the $\hat{pr}.general = 0.0199$ (See Equation 10[15]) is used.

5.1.2 Using data reported about alive and dead siblings

In this part, we incorporate dead siblings that were unaffectd or affected siblings in orginal-QEM analysis. However, the network size is still the size of number of living siblings, because probands can only be alive in order to be sampled and report about their (alive or dead) siblings' diabetes status.

Besides the variables needed in previous section, there are two additional variables needed: number of dead siblings (Col 510-511 on the public use file); number of dead siblings with diabetes (Col 512-513). The following are the frequency tables for the number of dead siblings, the dead siblings with diabetes, and the total number of siblings (alive+dead), respectively:

[14]RR is recurrence risk and RR is recurrence risk ratio.

[15]$\hat{pr}.general = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} w_i}$.

Table 14: 1976 NHIS Data: Frequencies for the number of dead siblings

dead siblings, No.	freq	dead siblings, No.	freq	dead siblings, No.	freq
0	82311	8	228	16	7
1	15464	9	139	17	3
2	5874	10	73	20	1
3	3007	11	60	22	3
4	1586	12	36	39	1
5	944	13	14	unknown	2504[16]
6	547	14	6		
7	362	15	8		

Table 15: 1976 NHIS Data: Frequencies for the number of affected dead siblings

affected (diabetic) dead siblings, No.	freq	affected dead siblings, No.	freq	affected dead siblings, No.	freq
0	109264	4	11	9	1
1	1216	5	3	unkown	2504
2	142	6	1		
3	35	7	1		

[16]There are 2504 probands reporting "unknown" for the number of their dead siblings. We treat them as missing values and they are not included in our study.

Table 16: 1976 NHIS Data: Frequencies for the number of total (alive+dead) siblings

total siblings, No.	freq	total siblings, No.	freq	total siblings, No.	freq
0	10817	11	1147	22	9
1	21975	12	760	23	3
2	21300	13	427	24	3
3	16409	14	194	25	1
4	11538	15	122	30	1
5	8224	16	60	41	1
6	5928	17	35	62	1
7	4468	18	17	70	1
8	3260	19	10	83	1
9	2301	20	17	unknown	2581[17]
10	1562	21	5		

In Table 16, we used 17 as the cut-off point and dropped all observations when the number of total siblings was larger than 17. After data editing, we obtain the parameter, recurrence risk and recurrence risk ratio estimates and their variance estimators given in Table 17.

[17]There are 2581 probands reporting at least one "unknown" for the number of their (alive or dead) siblings. We treat them as missing values and they are not included in our study.

Table 17: 1976 NHIS Data: Estimation results for original-QEM (alive and dead siblings)

	parameter estimate (var est)		RR & RRR estimate (var est)
δ	-3.80 (0.00031)	RR(original-QEM)	0.1541 (0.0000349)
γ	2.10 (0.00215)	RRR	7.07 (0.0581)

Again, the RR and RRR estimates in Table 17 agree with the nonparametric estimates in Graubard and Sirken (2013)[15]. Here, $\hat{pr}.general = 0.0218$[18] is used to calculate recurrence risk ratio estimate. Note the increase in the $\hat{pr}.general$ compared to the $\hat{pr}.general$ for alive siblings.

5.2 Relationship-EQEM

In this section, we will apply relationship-EQEM to 1976 NHIS diabetes data, hence diabetic information about proband's father and mother is also needed. Recall that when including the parents, there are three pairwise family relationships in our study: Father-Child, Mother-Child and Child-Child (i.e. Sib-Sib), and relationship-EQEM is given by,

[18] $\hat{pr}.general = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} w_i}$.

$$P(y_1, y_2|I_1, I_2) = \frac{e^{(\delta+\beta_1 I_1+\beta_2 I_2)y_1+(\delta+\beta_3 I_1+\beta_4 I_2)y_2+(\gamma+\beta_5 I_1+\beta_6 I_2)y_1 y_2}}{1 + e^{\delta+\beta_1 I_1+\beta_2 I_2} + e^{\delta+\beta_3 I_1+\beta_4 I_2} + e^{2\delta+\beta_1 I_1+\beta_2 I_2+\gamma+\beta_3 I_1+\beta_4 I_2+\beta_5 I_1+\beta_6 I_2}},$$

where $I_1 = 1, I_2 = 0, if Father-Child; I_1 = 0, I_2 = 1, if Mother-Child; I_1 = 0, I_2 = 0, if Child-Child$. Let y_1 be a binary variable for diabetic status for parent (Father or Mother) and y_2 is a binary variable for diabetic status for children (i.e., siblings of proband), and for Child-Child relationship, y_1 and y_2 are exchangeable.

There are eight variables needed for this analysis: proband diabetes status (Col 509 on the public use file); number of living siblings (Col 510-511); number of living siblings with diabetes (Col 512-513); sample weight for proband (Col 188-192); diabetes status for mother (Col 523); diabetes status for father (Col 525); living status for mother (Col 522); living status for father (Col 524). The following Table 18 and Table 19 are the frequencies for cross-classification of number of mothers with/without diabetes by living status and number of fathers with/without diabetes by living status, respectively:

Table 18: 1976 NHIS Data: Frequencies for number of Mothers (with/without diabetes, alive/dead)

	Diabetic Status			
Living Status	Not Affected	Affected	Unknown	Total
Alive	26526	3414	1173	31113
Dead	75334	4191	851	80376
Unknown	60	32	1597	1689
Total	101920	7637	3621	113178

Table 19: 1976 NHIS Data: Frequencies for number of Fathers (with/without diabetes, alive/dead)

	Diabetic Status			
Living Status	Not Affected	Affected	Unknown	Total
Alive	38816	2390	1611	42817
Dead	63797	2470	1474	68011
Unknown	86	3	2261	2350
Total	102699	5133	5346	113178

The frequencies in Tables 18 and 19 indicate the counts for reported families by sampled probands, since it is possible that two or more probands were sampled who are siblings from the same family. First we study only living siblings and their living parant(s), and next study both alive and dead siblings and parent(s) to calculate recurrence risk of diabetes among different family relationships, respectively. We will treat reported unknown diabetic status for the parents as not diabetic.

Additionally, we would like to emphasize the determination of "network size" here. According to the definition, network size is the number of possible reports for the familiy relationship pairs. In this section, our focus is three different relationship pairs: Father-Child; Mother-Child; Child-Child (Sib-Sib). For the first two types of family relationship pairs, the number of possible reports is the number of living siblings in the family, since all the interviewed probands were asked about the living status and diabetic status of their parents in 1976 NHIS. As for the third type of family relationship pair, Child-Child, the network size is also the number of living

siblings in the family. Therefore, the network size used in this section should be the total number of living siblings in the family, even when dead siblings and parents are also incorporated in the model.

5.2.1 Using data reported about living siblings and living parent(s)

In this section, we study recurrence risk of diabetes about living siblings and their living parent(s), but we don't require both parents to be alive. For families with living father and mother, we have Father-Child, Mother-Child and Child-Child relationship pairs; for families with living father but dead mother, we have Father-Child and Child-Child relationship pairs; for families with living mother but dead father, we have Mother-Child and Child-Child relationship pairs; while for families with both dead father and mother, we only have Child-Child relationship pairs. As in section 5.1, we delete information about probands who have more than 15 living siblings.

After data editing, we applied the relationship-EQEM to the 1976 NHIS data, obtaining the following results:

Table 20: 1976 NHIS Data: Estimation results for relationship-EQEM (living siblings and living parent(s))

	estimate(var est)		estimate(var est)		estimate(var est)
δ	-4.13(0.00026)	β_4	1.00(0.00045)	RR_{MC}	0.1095(0.0000149)
γ	2.43(0.00290)	β_5	-1.55(0.00443)	RR_{CM}	0.2431(0.0000582)
β_1	1.24(0.00079)	β_6	-1.40(0.00383)	RR_{CC}	0.1545(0.0000447)
β_2	1.96(0.00056)	RR_{FC}	0.0891(0.0000199)		
β_3	0.92(0.00039)	RR_{CF}	0.1191(0.0000343)		

Since there is ordering in relationship-EQEM, the recurrence risks of diabetes differ between conditioning on parent and conditioning on child. From Table 20, we notice that recurrence risk given parent's diabetes status is lower than the one given child's diabetes status. That is, $\hat{RR}_{FC}$ is less than $\hat{RR}_{CF}$ and $\hat{RR}_{MC}$ is less than $\hat{RR}_{CM}$. $\hat{RR}_{FC} > \hat{RR}_{CF}$ and $\hat{RR}_{MC} > \hat{RR}_{CM}$ indicate a stronger genetic component among parents and child when the child is affected with diabetes. Also recurrence risk of diabetes between mother and child is higher than father-child. This might be because children may be closer to their mothers than fathers and therefore, they are more likely to know if their mothers are diabetic than their fathers.

Next, we use Wald-tests to study recurrence risk differences among different family relationships.

Table 21: 1976 NHIS Data: Hypothesis testing results for relationship-EQEM (living siblings and living parent(s))

	H_0	DF	Test statistic	p-value
Wald-test.1	$H_0 : RR_{FC} = RR_{MC}$	1, 149	14.57	0.0002
Wald-test.2	$H_0 : RR_{FC} = RR_{CC}$	1, 149	108.44	< 0.00001
Wald-test.3	$H_0 : RR_{MC} = RR_{CC}$	1, 149	46.46	< 0.00001
Wald-test.4	$H_0 : RR_{CF} = RR_{CM}$	1, 149	204.43	< 0.00001
Wald-test.5	$H_0 : RR_{CF} = RR_{CC}$	1, 149	23.93	< 0.00001
Wald-test.6	$H_0 : RR_{CM} = RR_{CC}$	1, 149	94.99	< 0.00001
Wald-test.7	$H_0 : \begin{pmatrix} RR_{FC} - RR_{CC} \\ RR_{MC} - RR_{CC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	2, 148	53.85	< 0.00001
Wald-test.8	$H_0 : \begin{pmatrix} RR_{CF} - RR_{CC} \\ RR_{CM} - RR_{CC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	2, 148	101.56	< 0.00001

For Wald-test.1-Wald-test.8 in Table 21, all the p-values are less than 0.001 using an F-distribution which are highly statistically significant. Therefore, all H_0s are rejected and we obtain the following conclusions: recurrence risk differences exist between the Father-Child relationship and Mother-Child relationship; between the Father-Child relationship and Child-Child relationship; between the Mother-Child relationship and Child-Child relationship no matter whether conditioning on parent's diabetes status or conditioning on child's diabetes status, if we only consider living siblings and parent(s).

5.2.2 Using data reported about living siblings and living/dead parents

In this section, we study recurrence risk of diabetes about living siblings and their living or dead parent(s), and we also choose 15 as the cut-off point as section 5.1. Table 22 shows the model parameter estimates and recurrence risk estimates for each relationship.

Table 22: 1976 NHIS Data: Estimation results for relationship-EQEM(living siblings and alive/dead parent(s))

	estimate(var est)		estimate(var est)		estimate(var est)
δ	-4.13(0.00026)	β_4	0.03(0.00028)	RR_{MC}	0.0690(0.0000047)
γ	2.43(0.00290)	β_5	-1.30(0.00313)	RR_{CM}	0.2356(0.0000395)
β_1	1.06(0.00049)	β_6	-0.93(0.00351)	RR_{CC}	0.1545(0.0000447)
β_2	1.46(0.00037)	RR_{FC}	0.0543(0.0000051)		
β_3	0.15(0.00024)	RR_{CF}	0.1253(0.0000248)		

Next, we use Wald tests to study recurrence risk differences among different family relationships when studying on living siblings and their living and dead parent(s).

Table 23: 1976 NHIS Data: Hypothesis testing results for relationship-EQEM (living siblings and living parent(s))

	H_0	DF	Test statistic	p-value
Wald-test.1	$H_0 : RR_{FC} = RR_{MC}$	1, 149	28.94	< 0.00001
Wald-test.2	$H_0 : RR_{FC} = RR_{CC}$	1, 149	283.26	< 0.00001
Wald-test.3	$H_0 : RR_{MC} = RR_{CC}$	1, 149	180.75	< 0.00001
Wald-test.4	$H_0 : RR_{CF} = RR_{CM}$	1, 149	222.10	< 0.00001
Wald-test.5	$H_0 : RR_{CF} = RR_{CC}$	1, 149	19.56	< 0.00001
Wald-test.6	$H_0 : RR_{CM} = RR_{CC}$	1, 149	110.25	< 0.00001
Wald-test.7	$H_0 : \begin{pmatrix} RR_{FC} - RR_{CC} \\ RR_{MC} - RR_{CC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	2, 148	148.94	< 0.00001
Wald-test.8	$H_0 : \begin{pmatrix} RR_{CF} - RR_{CC} \\ RR_{CM} - RR_{CC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	2, 148	111.38	< 0.00001

For Wald-test.1-Wald-test.8 in Table 23, all the p-values are less than 0.0005 using an F-distribution which are highly statistically significant. Therefore, all H_0s are rejected and we obtain the same conclusions as Section 5.2.1 when considering living siblings and their living/dead parents.

5.2.3 Using data reported about alive/dead siblings and alive/dead parents

In this section, we consider diabetes data reported for both living and dead siblings and their living or dead parents. Note that reports of greater than 17 total siblings

and parents are deleted in our analysis because of a concern that there may be inaccurate reporting by the proband. Also, the network size is the number of living siblings in the family. Table 24 shows the model parameter estimates and recurrence risk estimates for each relationship.

Table 24: 1976 NHIS Data: Estimation results for relationship-EQEM(living/dead siblings and alive/dead parent(s))

	estimate(var est)		estimate(var est)		estimate(var est)
δ	-3.80(0.00031)	β_4	-0.20(0.00023)	RR_{MC}	0.0715(0.0000057)
γ	2.10(0.00215)	β_5	-0.97(0.00386)	RR_{CM}	0.2328(0.0000454)
β_1	0.69(0.00049)	β_6	-0.66(0.00323)	RR_{CC}	0.1541(0.0000349)
β_2	1.17(0.00052)	RR_{FC}	0.0594(0.0000065)		
β_3	-0.09(0.00019)	RR_{CF}	0.1207(0.0000258)		

Next, we use Wald tests to study recurrence risk differences among different family relationships when studying both living and dead siblings and parents.

Table 25: 1976 NHIS Data: Hypothesis testing results for relationship-EQEM (living/dead siblings and parents)

	H_0	DF	Test statistic	p-value
Wald-test.1	$H_0 : RR_{FC} = RR_{MC}$	1, 149	15.35	0.00014
Wald-test.2	$H_0 : RR_{FC} = RR_{CC}$	1, 149	265.80	< 0.00001
Wald-test.3	$H_0 : RR_{MC} = RR_{CC}$	1, 149	212.40	< 0.00001
Wald-test.4	$H_0 : RR_{CF} = RR_{CM}$	1, 149	205.06	< 0.00001
Wald-test.5	$H_0 : RR_{CF} = RR_{CC}$	1, 149	23.37	< 0.00001
Wald-test.6	$H_0 : RR_{CM} = RR_{CC}$	1, 149	94.08	< 0.00001
Wald-test.7	$H_0 : \begin{pmatrix} RR_{FC} - RR_{CC} \\ RR_{MC} - RR_{CC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	2, 148	132.73	< 0.00001
Wald-test.8	$H_0 : \begin{pmatrix} RR_{CF} - RR_{CC} \\ RR_{CM} - RR_{CC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	2, 148	102.30	< 0.00001

For Wald-test.1-Wald-test.8 in Table 25, all the p-values are less than 0.0005 using an F-distribution which are highly statistically significant. Therefore, all H_0s are rejected and we obtain the same basic conclusions as Section 5.2.1 if we consider living/dead siblings and their living/dead parents.

5.3 Race-EQEM

There are four parameters in the race-EQEM: δ_W, γ_W, δ_O and γ_O indicating parameters for whites and other races, respectively. The variable for race is located in Column 51 in 1976 NHIS diabetes data public use file, and value 1 means whites, while value 2 denotes other races. Similary as in previous sections, we will focus on living siblings only, and both alive and dead siblings respectively to estimate parameters, recurrence

risk and recurrence risk ratio. The following is the unweighted frequencies for race:

Table 26: 1976 NHIS Data: Frequencies for race

race	frequency
whites	85653
other races	11858

5.3.1 Using data reported about only living siblings

As in Section 5.1, we deleted observations where greater than 15 siblings are reported. The results for the parameter and recurrence risks estimation are shown in following table:

Table 27: 1976 NHIS Data: Estimation results for race-EQEM (only living siblings)

	estimate(var est)		estimate(var est)		estimate(var est)
δ_W	-4.10(0.00028)	γ_O	2.61(0.01879)	RRR_W	7.75(0.0950)
γ_W	2.40(0.00273)	RR_W	0.1544(0.0000449)	RRR_O	7.70(0.7320)
δ_O	-4.30(0.00195)	RR_O	0.1552(0.000298)		

From Table 27, we see that the recurrence risk of diabetes for whites is slightly lower than for other races, while the recurrence risk ratio is slightly higher for whites. This is due to the prevelence of diabetes being lower for whites than other races. However, difference between two kinds of races are small. We expect that the variance for RR and RRR is smaller for whites, since their sample size is much larger than

for other races. Note that $\hat{pr}.general$[19] is used to calculate prevelance of diabetes. The results for $\hat{RR}_W$, $\hat{RR}_O$, $R\hat{R}R_W$ and $R\hat{R}R_O$ agree with the model-free method (Graubard and Sirken, 2013)[15].

Next, we will conduct Wald-tests and Quasi-score tests to study the differences in parameters and recurrence risks between whites and other races.

Wald-test.1 Let $H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Hence, the test statistic $F_W = 9.55$ with $DF_1 = 2$ and $DF_2 = 148$. Thus, the p-value is 0.00013 using an F-distribution and H_0 should be rejected. Thus we conclude that there is a difference in the parameter (δ_W, δ_O) and (γ_W, γ_O). Next, the Quasi-score test is applied to test H_0.

Quasi-score test $H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. After calculation, we obtain $\hat{S}(\widetilde{\Phi}) = (261202.66, 19634.22, -261202.66, -19634.22)^T$,

$$\widetilde{I}_S = \begin{pmatrix} 9897288.7 & 1342739.6 & 0 & 0 \\ 1342739.6 & 682006.7 & 0 & 0 \\ 0 & 0 & 1928629.2 & 261652.1 \\ 0 & 0 & 261652.1 & 132898.8 \end{pmatrix},$$

$$\widetilde{V}_S = \begin{pmatrix} 4.56 \times 10^9 & 7.41 \times 10^8 & -4.56 \times 10^9 & -7.41 \times 10^8 \\ 7.41 \times 10^8 & 2.03 \times 10^8 & -7.41 \times 10^8 & -2.03 \times 10^8 \\ -4.56 \times 10^9 & -7.41 \times 10^8 & 4.56 \times 10^9 & 7.41 \times 10^8 \\ -7.41 \times 10^8 & -2.03 \times 10^8 & 7.41 \times 10^8 & 2.03 \times 10^8 \end{pmatrix}, \text{ so } F_S = 21.325$$

with degrees of freedom $DF_1 = 2$ and $DF_2 = 148$. Hence, the p-value is 0.00005 using an F-distribution and H_0 is rejected. Thus, we obtain the same conclusion as

[19] $\hat{pr}.general = \frac{\sum_{i=1}^n \frac{w_i}{s_i} a_i}{\sum_{i=1}^n \frac{w_i}{s_i} s_i} = \frac{\sum_{i=1}^n \frac{w_i}{s_i} a_i}{\sum_{i=1}^n w_i}$.

Wald-test. The parameter differences between the two race groups implies that the model parameters differ between the races but does not imply that the RR_W and RR_O differ. Thus, the following test is designed to address differences in RR_W and RR_O.

Wald-test.2 Let $H_0 : RR_W = RR_O$. Then, the test statistic is 0.00248 with degrees of freedom $DF_1 = 1$, $DF_2 = 149$. Hence, the p-value is 0.9603 using an F-distribution and H_0 cannot be rejected. Therefore, we conclude there is no difference for sibling recurrence risk between whites and other races when we restrict the analysis to only living siblings, even though we cannot accept that the parameter estimates are the same between different race groups.

5.3.2 Using data reported about alive and Dead siblings

As in Section 5.1 and we deleted observations where more than 17 siblings (living or dead) are reported. The results for the parameter and recurrence risks estimation are shown in following table:

Table 28: 1976 NHIS Data: Estimation results for race-EQEM (living/dead siblings)

	estimate(var est)		estimate(var est)		estimate(var est)
δ_W	-3.77(0.00030)	γ_O	2.03(0.01561)	RRR_W	7.15(0.0673)
γ_W	2.10(0.00241)	RR_W	0.1583(0.0000419)	RRR_O	6.48(0.4319)
δ_O	-3.96(0.00265)	RR_O	0.1271(0.000167)		

Wald-test.1 Similarly, let $H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Then the test statistic is 8.28 with $DF_1 = 2$ and $DF_2 = 148$. Hence, the p-value is 0.00039 using an F-distribution. Thus H_0 is rejected and we conclude that a significant difference exists between (δ_W, δ_O) and (γ_W, γ_O).

Wald-test.2 Let $H_0 : RR_W = RR_O$. Then, the test statistic is 4.7542 with $DF_1 = 1$, $DF_2 = 298 - 149 = 149$. Hence, the p-value is 0.0308 using an F-distribution. Thus H_0 is rejected and we conclude that a significant difference exists between RR_W and RR_O if we consider both living and dead siblings.

Comparing estimation results using data reported for only living siblings with using both living and dead siblings, we can see that the sibling recurrence risk for other races is smaller, and the difference of recurrence risk between whites and other races becomes larger after including dead siblings in the model. This indicates that the recurrence risk is smaller among dead siblings for other races. This could be explained by larger reporting error of diabetes for dead siblings than for living siblings and this reporting error is more pronounced among the other race group.

5.4 Addressing an age effect

In this section, propensity score adjustment will be applied using propensity score weighting to remove potential age effects in studying the relationship between race and recurrence risk of diabetes. In this section, we will only study recurrence risk of diabetes among living siblings.

In Section 5.3, we estimated that sibling recurrence risk has a weak relationship with race, $RR_W = 0.1544$ and $RR_O = 0.1552$. Because we found out that whites are younger than other races (the correlation coefficient between proband's age and race in NHIS data is -0.0543) and the prevalence of diabetes increases with increasing age (the correlation coefficient between proband's age and diabetes status in NHIS data is 0.1705), age can confound the estimation of recurrence risk of diabetes. Thus, we conduct propensity score weighting to reduce the age effect in our analysis of race effect on recurrence risk of diabetes.

Unfortunately, age is not reported for the siblings in the 1976 NHIS. Therefore, to illustrate the use of propensity score weighting, we simulate ages for siblings based on the reported age of the proband. We simulate ages for siblings according to a uniform disbribution in the interval (proband's age (yr)-5, proband's age (yr)+5). If the proband's age (yr)-5 is less than zero, then we use a uniform(0,10). We assume here that the range of ages among siblings are within 10 years.

After generating ages for each sib-pair, we use the average age for each pair of siblings and then calculate propensity score weights based on ATE method (the details have been shown in Section 3.4). Note that the "new" weight is the product of propensity score weight and sample weight (which is the sample weight for proband divided by network size). The following are weighted density plots for average age between whites and other races before and after propensity score weighting, respectively. We can see the age differences has been reduced after the propensity score weighting even though the age differences between whites and other races are small before adjustment.

Figure 11: 1976 NHIS Data: Age Density before Propensity Score Analysis (ATE)

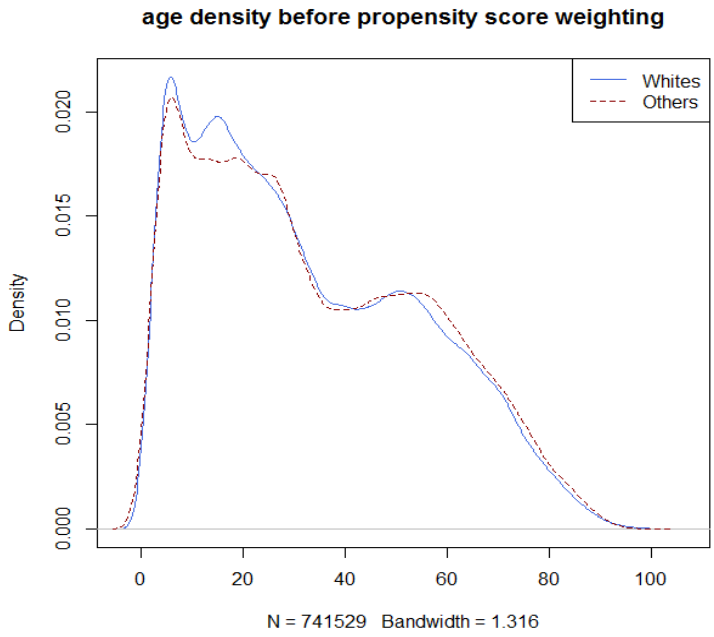
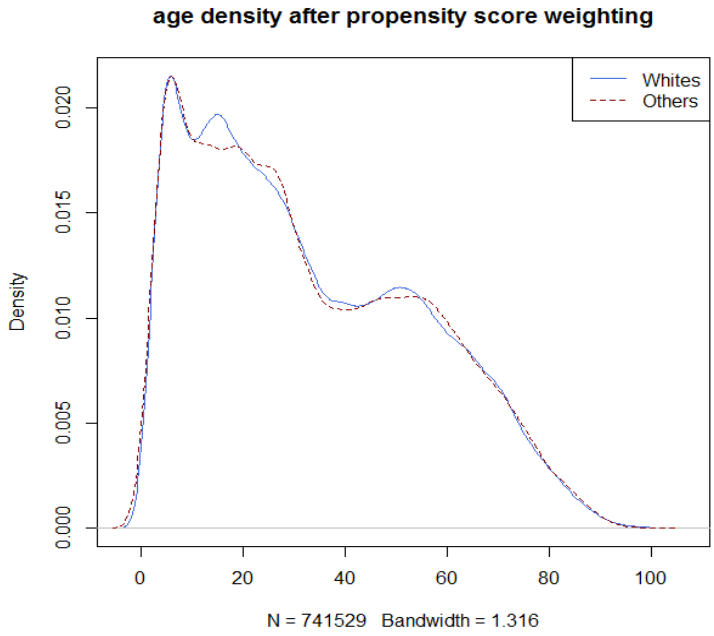


Figure 12: 1976 NHIS Data: Age Density after Propensity Score Analysis (ATE)



We also computed the mean value of paired average age: before propensity score weighting, the mean of paired average age for whites is 32.777, while it is 33.614 for

other races; after propensity score weighting, the mean of paired average age for the whites is 32.912, while it is 32.905 for other races. Therefore, after propensity score weighting, age differences are reduced between whites and other races. The following table is the estimation results after propensity score adjustment for age:

Table 29: 1976 NHIS Data: Estimation results for race-EQEM(only living siblings, after propensity score weighting)

	estimate(var est)		estimate(var est)		estimate(var est)
δ_W	-4.10(0.00028)	γ_O	2.60(0.01882)	RRR_W	7.75(0.0949)
γ_W	2.40(0.00273)	RR_W	0.1544(0.0000449)	RRR_O	7.71(0.7278)
δ_O	-4.30(0.00199)	RR_O	0.1552(0.000297)		

Therefore, estimation of recurrence risks and recurrence risk ratios by race categories after propensity score weighting is similar to estimation without propensity score weighting, although the estimation results are more accurate after propensity score weighting because it reduces the confounding effect of age.

Below, we conduct Wald-tests and Quasi-score tests using the propensity score weighting to study differences between whites and other races.

Wald-test.1 Let $H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. The test statistic is equal to 9.29 with $DF_1 = 2$, $DF_2 = 148$. Since the p-value is 0.00016 using an F-distribution, H_0 is rejected and we conclude that differences between the parameters (δ_W, δ_O) and (γ_W, γ_O) exist after age adjustment.

Quasi-score test Let $H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Similar to Section 5.3 before age adjustment, we have $H = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$,

$\hat{S}(\widetilde{\Phi}) = (954366.18, 71684.88, -954366.18, -71684.88)^T$,

$\widetilde{I}_S = \begin{pmatrix} 11120447 & 1509917 & 0 & 0 \\ 1509917 & 766177 & 0 & 0 \\ 0 & 0 & 11198327 & 1520492 \\ 0 & 0 & 1520492 & 771543 \end{pmatrix}$, and

$\widetilde{V}_S = \begin{pmatrix} 6.10 \times 10^{10} & 9.99 \times 10^9 & -6.10 \times 10^{10} & -9.99 \times 10^9 \\ 9.99 \times 10^9 & 2.76 \times 10^9 & -9.99 \times 10^9 & -2.76 \times 10^9 \\ -6.10 \times 10^{10} & -9.99 \times 10^9 & 6.10 \times 10^{10} & 9.99 \times 10^9 \\ -9.99 \times 10^9 & -2.76 \times 10^9 & 9.99 \times 10^9 & 2.76 \times 10^9 \end{pmatrix}$. Therefore, the test statistic $F_S = 21.324$ with $DF_1 = 2$ and $DF_2 = 148$. Hence, the p-value is 0.00005 using an F-distribution and H_0 is rejected, and we conclude that there are differences between the parameters (δ_W, δ_O) and (γ_W, γ_O) after adjustment for age. Because parameter differences between race groups does not imply the differences in the recurrence risk between whites and other races, we conduct a Wald-test to test differences in recurrence risk between whites and other races.

Wald-test.2 Let $H_0 : RR_W = RR_O$. The test statistic is 0.000026 with $DF_1 = 1$, $DF_2 = 149$. Hence, the p-value is 0.9595 using an F-distribution and H_0 is not rejected so we conclude there is no difference for sibling recurrence risk between whites and other races after age adjustment.

Chapter 6 Simulation study

In this chapter we present a simulation study where we generate large finite populations of families of siblings with their disease statuses based on a bivariate density conditional on their parents' disease statuses. We generate families of sibsize $1-7$ in our simulation study.

The simulation study is mostly based on parameter values from the 1976 NHIS diabetes data. The estimators used for 1976 NHIS to obtain values for parameters needed in the simulation study are given below:

(1) Prevalence of disease for males,

$$p^{(m)} = \frac{\sum_{i=1}^{n} w_i I_i\{male\} I_i\{D=1\}}{\sum_{i=1}^{n} w_i I_i\{male\}},$$

where $I_i\{.\}$ is an indicator variable that equals 1 if $\{.\}$ is satified and 0 otherwise. This estimation is based on only proband information, $D=1$ means proband is diabetic. Additionally, w_i denotes sample weight of i^{th} proband. n is total number of observations in 1976 NHIS.

(2) Prevalence of disease for females,

$$p^{(f)} = \frac{\sum_{i=1}^{n} w_i I_i\{female\} I_i\{D=1\}}{\sum_{i=1}^{n} w_i I_i\{female\}}.$$

(3) Estimate of numbers of families with sibsize k,

$$NS_{ksib} = \sum_{i=1}^{n} \frac{w_i I_i\{sibsize\ k\}}{s_i},$$

where $k = 1, 2, ..., 7$ and s_i is alive sibsize of the family of i^{th} proband. Note that this is a network estimator.

(4) Proportion of families with sibsize k, $p_{ksib} = \frac{NS_{ksib}}{NS_{total}}$, where $NS_{total} = \sum_{k=1}^{7} NS_{ksib}$.

(5) Average age of probands,

$$\bar{a} = \frac{\sum_{i=1}^{n} w_i a_i}{\sum_{i=1}^{n} w_i},$$

where a_i is age of i^{th} proband.

(6) Proportion of children with diabetes conditional on parents' disease status:

a. Proportion of probands where both mother and father are diabetic,

$$p_{11} = \frac{\sum_{i=1}^{n} w_i I_i\{DM = 1, DF = 1, D = 1\}}{\sum_{i=1}^{n} w_i I_i\{DM = 1, DF = 1\}}.$$

b. Proportion of probands where only mother is diabetic,

$$p_{10} = \frac{\sum_{i=1}^{n} w_i I_i\{DM = 1, DF = 0, D = 1\}}{\sum_{i=1}^{n} w_i I_i\{DM = 1, DF = 0\}}.$$

c. Proportion of probands where only father is diabetic,

$$p_{01} = \frac{\sum_{i=1}^{n} w_i I_i\{DM = 0, DF = 1, D = 1\}}{\sum_{i=1}^{n} w_i I_i\{DM = 0, DF = 1\}}.$$

d. Proportion of probands where neither mother nor father is diabetic,

$$p_{00} = \frac{\sum_{i=1}^{n} w_i I_i\{DM = 0, DF = 0, D = 1\}}{\sum_{i=1}^{n} w_i I_i\{DM = 0, DF = 0\}}.$$

Here $DM = 1$ means mother is diabetic and similarly, $DF = 1$ means father is diabetic for above estimators a-d.

The above estimates are obtained for the different races: whites and other races. We use the subscripts W and O to represent whites and other races in the following simulation study.

6.1 Simulation steps

In this section, we introduce simulation steps. Note that we will adjust for age effects using propensity score weighting before analyzing the relationship between recurrence risk of diabetes and race.

Step 1 Suppose there are N_f families with varying family sizes in the finite populations for the simulation study. Generate $0.65 * N_f$ pairs of white parents' diabetic status with prevalence of disease (Note that we set 65% families are

whites based on current U.S. population values, not from 1976 NHIS, since there were few families with other races in 1976), $p_W^{(m)}$ for white father and $p_W^{(f)}$ for white mother estimated from 1976 NHIS diabetes data, and generate $0.35 * N_f$ pairs of other race parents' diabetic status with prevalence of disease, $p_O^{(m)}$ for father with other races and $p_O^{(f)}$ for mother with other races. To repear here we assume that 65% of the finite population is whites, while the other 35% is other race.

Step 2 Determine the number of families with sibsize $1-7$ respectively, using multinomial distribution:

N_{1sibW}, N_{2sibW},...,N_{7sibW} follows $Multinomial(0.65{*}N_f, p_{1sibW}, p_{2sibW}, ..., p_{7sibW})$, N_{1sibO}, N_{2sibO},...,N_{7sibO} follows $Multinomial(0.35 * N_f, p_{1sibO}, p_{2sibO}, ..., p_{7sibO})$ for whites and other races respectively, where the probabilities are estimated from 1976 NHIS diabetes data. The values of $p_{1sibW} - p_{7sibW}$, $p_{1sibO} - p_{7sibO}$ are shown at the beginning of Section 6.3.

Step 3 Use binary distribution to generate the disease status of children according to empirical conditional proportions $(p_{11W}, p_{10W}, p_{01W}, p_{00W}, p_{11O}, p_{10O}, p_{01O}, p_{00O})$. Note that p_{11} denotes the probability of child with diabetes when both mother and father are diabetic, p_{10} is when only mother is diabetic, p_{01} is when only father is diabetic, and p_{00} denotes the probability of child with diabetes when neither of father and mother are diabetic, and sub-script W and O denote whites and other races. Repeat the same process to generate other children's disease status for each pair of parents if sibsize is more than one. Then we have two parents and k child/children($k = 1, 2, ..., 7$) for each family. Note that we are assuming that disease status is conditionally independent for the siblings, given their parents' disease status. That is $P(Dsib1 = 1, Dsib2 = 1, ..., Dsibk = 1|DF, DM) = P(Dsib1 = 1|DF, DM) \times P(Dsib2 = 1|DF, DM) \times ... \times P(Dsibk =$

$1|DF, DM)$, $k = 1, 2, ..., 7$. The values of p_{11W}, p_{10W}, p_{01W}, p_{00W}, p_{11O}, p_{10O}, p_{01O}, p_{00O} are shown in Section 6.3.

Step 4 Obtain estimated average age for proband from 1976 NHIS data, stratified by race: $\bar{a}_W$ and $\bar{a}_O$. Use these values as means of normal distributions for simulating ages for probands in each family (We assume the standard deviations for the normal distribution are 1 for both whites and other races). Then generate ages for the siblings of each proband using uniform distribution with minimal value (proband's age (yr)-5) and maximum value (proband's age (yr)$+5$). For all simulated ages, if the age is less than 1, then we set it to be 1; while if simulated age is larger than 90, then we set it to be 90.

Step 5 The siblings (i.e., children of the parents) form a finite population. Hence, there is only one generation in the finite population. We will use only one finite population for our simulations of relationship-EQEM and race-EQEM studies.

Step 6 Assign strata and PSUs to the finite population of siblings by randomly assigning the finite population to 100 strata and within each stratum, assign 25 PSUs allowing for positive intra-cluster coefficient. The details for strata and PSUs assignment will be described in Section 6.2.

Step 7 Adjust for age effects before estimation of recurrence risk. The adjustment procedure is given in Chapter 3. Note that the weight used in estimation is the product of sample weight and propensity score weight after age effect adjustment.

Step 8 Estimate values of recurrence risk under relationship-EQEM and race-EQEM based on the entire finite population and treat these values as true values.

Step 9 We use two different sample designs to select samples in the simulation study. First is a two-stage cluster SRS: randomly sample n_{s1} PSUs per stratum, and

then sample n_{s2} people per sampled PSU. Hence there are $100 \times n_{s1} \times n_{s2}$ sampled individuals in total (recall there are 100 strata). Because the sample weighting is the same for each sampled person, this is a noninformative sample weighting design. The second sample design has informative sample weights. We group individuals in the selected PSUs into two strata: one consists of individuals with diabetes; the other consists of individuals without diabetes. Let the probability for sampling a diabetic individual be denoted as pr_d, while the probability for sampling an non-diabetec individual is pr_n. Rounding the probabilities of selection is needed if the sample size is not an integer. It is possible that there are no diabetic individuals in a sampled PSU since the probability of being diabetic is small.

Step 10 Repeat Step 9 1000 times, average the parameter estimates among 1000 simulations and make comparisons between the true values based on the finite population and the parameter estimates obtained from 1000 samples. Sample design-based variance estimates averaged over the 1000 samples are compared with empirical variances obtained over 1000 simulations.

6.2 Strata and PSUs assignment

Now, we give further details for Step 6 in the simulation. First, the finite population of individuals are assigned randomly to 100 strata. We partition individuals within each stratum into 25 PSUs so that there is intra-cluster correlation with respect to an individual-level variable that is related to being diabetic. This will induce intra-cluster correlation for diabetic status of individuals within each PSU.

In the simulation study, we let the individual-level variable follow a standard normal distribution with zero mean and standard deviation of one. Individuals in the

population who are non-diabetic have means $\mu_{nW} = \mu_{nO} = 0$ for the individual-level variable for whites and other races, respectively. For individuals in the population who are diabetic, the individual-level variable will have means μ_{dW} and μ_{dO} for whites and other races, respectively. The μ_{dW} and μ_{dO} will be determined to induce intra-cluster correlation in the PSUs by partitioning individuals with similar individual-level variable values into the same PSU. Note that the assignment of strata is independent of the normally distributed individual-level variable.

Next we determine the values of μ_{dW} and μ_{dO} to set the intra-cluster correlation to be a specified value. Intra-cluster correlation coefficient (ICC) of a variable describes how similar values of the variable are in the same cluster compared to values of the variable between clusters. The formula of ICC is given by,

$$ICC = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2},$$

where σ_A^2 and σ_E^2 are the population model-based variance components, σ_A^2 is the between cluster variance component and σ_E^2 is the within cluster variance component. These variance components for the finite population are computed as $\hat{\sigma}_E^2 = \sum_{i=1}^{K} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.})^2 / \sum_{i=1}^{K} (M_i - 1)$, $\hat{\sigma}_A^2 = (S_B^2 - \hat{\sigma}_E^2)^2 / M_0$, $S_B^2 = \sum_{i=1}^{K} M_i (\bar{y}_{i.} - \bar{y}_{..})^2$ and $M_0 = \frac{1}{K-1}(N - \sum_{i=1}^{K} M_i^2 / N)$, where K denotes the numbers of PSUs in the finite population, M_i is the numbers of individuals in the i^{th} PSU, N is the population size. y_{ij} is the diabetes status for the j^{th} individual in the i^{th} PSU, $\bar{y}_{i.}$ denotes the mean value of diabetes status for i^{th} PSU, and $\bar{y}_{..}$ is the overall mean for the whole finite population[13].

In our simulation settings, we set $ICC = 0.05$, which is pre-determined. We tried

different values of μ_{dW} and μ_{dO} and then to calculate the corresponding value of ICC. And finally determine values of μ_{dW} and μ_{dO} to achieve the pre-determined value of $ICC = 0.05$. The values of μ_{dW} and μ_{dO} are shown in Table 30 and Table 31.

6.3 Simulation results

According to simulation steps, we conduct simulation studies under relationship-EQEM and race-EQEM, and both SRS and informative sampling methods are used to select samples. Here are population values that we use in our simulation that are estimated from 1976 NHIS diabetes data:

(1) prevalence of diabetes for females and males: $p_W^{(f)} = 0.0232$, $p_W^{(m)} = 0.0196$, $p_O^{(f)} = 0.0301$, $p_O^{(m)} = 0.0220$.

(2) proportion of families with sibsize $1-7$ (whites and other races): $p_{1sibW} = 0.2524$, $p_{2sibW} = 0.2931$, $p_{3sibW} = 0.1987$, $p_{4sibW} = 0.1184$, $p_{5sibW} = 0.0687$, $p_{6sibW} = 0.0420$, $p_{7sibW} = 0.0267$; $p_{1sibO} = 0.3561$, $p_{2sibO} = 0.2205$, $p_{3sibO} = 0.1505$, $p_{4sibO} = 0.1063$, $p_{5sibO} = 0.0758$, $p_{6sibO} = 0.0536$, $p_{7sibO} = 0.0372$.

(3) proportion of children with diabetes (whites and other races): $p_{11W} = 0.1146$, $p_{10W} = 0.0648$, $p_{01W} = 0.0471$, $p_{00W} = 0.0166$; $p_{11O} = 0.1150$, $p_{10O} = 0.0774$, $p_{01O} = 0.0532$, $p_{00O} = 0.0216$.

(4) average age of proband (whites and other races): $\bar{a}_W = 33$, $\bar{a}_O = 28$.

Additionally, the prevalence of disease used for recurrence risk ratio calculation is based on $\hat{pr}.general$ shown in Equation 9. Also, instead of generating one fi-

nite population and regenerating samples, we can also implement a super-population simulation, that is, we regenerate the whole population 1000 times, and for each regeneration we select one sample from the population each time. But due to the large amount of time, we only run simulations from a single finite population in our research.

At the end of this chapter, Type I error checking and power analysis are investigated for hypothesis testings which are shown in Chapter 5. Note that we repeat the simulation 1000 times to compute the Type I error.

Table 30: Simulation: Simulation Result for Relationship-EQEM

population size and sample size		parameter estimate based on finite population	estimate based on SRS samples (var. est)	estimate based on informative samples (var. est)	empirical variance 1.SRS 2.Informative
	δ	-3.8708	-3.8701	-3.8703	(0.00175)
			(0.00188)	(0.00180)	(0.00173)
	γ	0.1403	0.1329	0.1382	(0.0213)
			(0.0208)	(0.0139)	(0.0133)
	β_1	-0.0127	-0.0132	-0.0129	(0.0038)
			(0.0040)	(0.00264)	(0.00259)
$N_f = 380000$	β_2	0.1658	0.1618	0.1642	(0.0033)
$N_{pop} = 1015642$			(0.0035)	(0.0023)	(0.0021)
$n_{s1} = 2$	β_3	-0.0369	-0.0388	-0.0372	(0.00090)
$n_{s2} = 100(SRS)$			(0.00094)	(0.00073)	(0.00067)
$pr_d = 50\%$	β_4	-0.0744	-0.0759	-0.0746	(0.00095)
			(0.00098)	(0.00078)	(0.00070)
$pr_{nw} = 20\%$	β_5	0.8400	0.8523	0.8471	(0.0366)
$pr_{no} = 25\%$			(0.0373)	(0.0215)	(0.0211)
$\mu_{dw} = 1.8$	β_6	1.2570	1.2687	1.2519	(0.0275)
$\mu_{do} = 1.2$			(0.0284)	(0.0176)	(0.0157)
	RR_{FC}	0.0508	0.05095	0.05092	(0.000047)
			(0.000048)	(0.000027)	(0.000026)
	RR_{CF}	0.05120	0.0522	0.05123	(0.000045)
			(0.000048)	(0.000024)	(0.000022)
	RR_{MC}	0.07256	0.07250	0.07253	(0.000057)
			(0.000058)	(0.0000331)	(0.0000327)
	RR_{CM}	0.0905	0.0902	0.0903	(0.000083)
			(0.000082)	(0.000041)	(0.000043)
	RR_{CC}	0.0234	0.02329	0.02336	(0.0000112)
			(0.0000111)	(0.0000074)	(0.0000071)

Table 31: Simulation: Simulation Result for Race-EQEM

population size and sample size		parameter estimate based on finite population	estimate based on SRS samples (var. est)	estimate based on informative samples (var. est)	empirical variance 1.SRS 2.Informative
	δ_W	-3.9738	-3.9732	-3.9740	(0.00299)
			(0.0031)	(0.0030)	(0.00288)
$N_f = 380000$	γ_W	0.0741	0.0627	0.0706	(0.0426)
$N_{pop} = 1015642$			(0.0435)	(0.0299)	(0.0293)
$n_{s1} = 2$	δ_O	-3.7081	-3.7086	-3.7084	(0.0022)
$n_{s2} = 100(SRS)$			(0.0023)	(0.0020)	(0.0018)
$pr_d = 50\%$	γ_O	0.1918	0.1796	0.1889	(0.0426)
			(0.0433)	(0.0276)	(0.0258)
$pr_{nw} = 20\%$	RR_W	0.01985	0.0196	0.01986	(0.0000155)
$pr_{no} = 25\%$			(0.0000159)	(0.0000112)	(0.0000110)
$\mu_{dw} = 1.8$	RR_O	0.02885	0.02880	0.02886	(0.0000341)
$\mu_{do} = 1.2$			(0.0000345)	(0.000021)	(0.000020)
	RRR_W	1.2426	1.2269	1.2434	(0.0578)
			(0.0594)	(0.0451)	(0.0428)
	RRR_O	1.2024	1.2004	1.2020	(0.0568)
			(0.0572)	(0.0369)	(0.0341)

[19]In Table 30 and Table 31, N_f denotes the total family numbers in the finite population, N_{pop} denotes the population size (individuals), n_{s1} is the number of PSUs sampled from each stratum, n_{s2} is the number of individuals sampled from each sampled PSU when SRS is used, pr_d is the probability for sampling a diabetic individual, pr_{nw} is the probability for sampling a non-diabetic whites, pr_{no} is the probability for sampling a non-diabetic individual with other races. μ_{dw} and μ_{do} is the mean of normal distribution for the individual-level variable used for intra-cluster correlation with diabetes for whites and other races respectively.Note that the means for non-diabetic normal distributions are $\mu_{nw} = \mu_{no} = 0$.

Table 32: Simulation: Type I error results

type	null hypothesis	sample method	Type I error
Wald test	$H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} -0.2262 \\ -0.0169 \end{pmatrix}$	SRS	0.044
Wald test	$H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} -0.2262 \\ -0.0169 \end{pmatrix}$	Informative	0.056
Quasi-score test	$H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} -0.2262 \\ -0.0169 \end{pmatrix}$	SRS	0.063
Quasi-score test	$H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} -0.2262 \\ -0.0169 \end{pmatrix}$	Informative	0.061
Wald test	$H_0 : RR_W - RR_O = -0.0031$	SRS	0.049
Wald test	$H_0 : RR_W - RR_O = -0.0031$	Informative	0.054
Wald test	$H_0 : RR_{FC} - RR_{MC} = -0.0091$	SRS	0.058
Wald test	$H_0 : RR_{FC} - RR_{MC} = -0.0091$	Informative	0.045
Wald test	$H_0 : RR_{FC} - RR_{CC} = 0.0138$	SRS	0.054
Wald test	$H_0 : RR_{FC} - RR_{CC} = 0.0138$	Informative	0.045

As for power analysis, we use the same set up as for the Type I error study. Note that both SRS and informative sampling methods are used and replication times are 100 for power analysis. However, except SRS with ICC and informative sampling approaches used for power analysis, there is another sampling approach used: SRS with $ICC = 0$, which means strata and clusters are assigned to the population randomly without intra-cluster coefficient.

$$\text{Let } H_0 : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} -0.2262 \\ -0.0169 \end{pmatrix}$$

$$H_a : \begin{pmatrix} \delta_W - \delta_O \\ \gamma_W - \gamma_O \end{pmatrix} = \begin{pmatrix} -0.2262 - \Delta \\ -0.0169 - \Delta \end{pmatrix}$$

Figure 13: Simluation: Power Analysis (1.1). Wald test and Quasi-score test

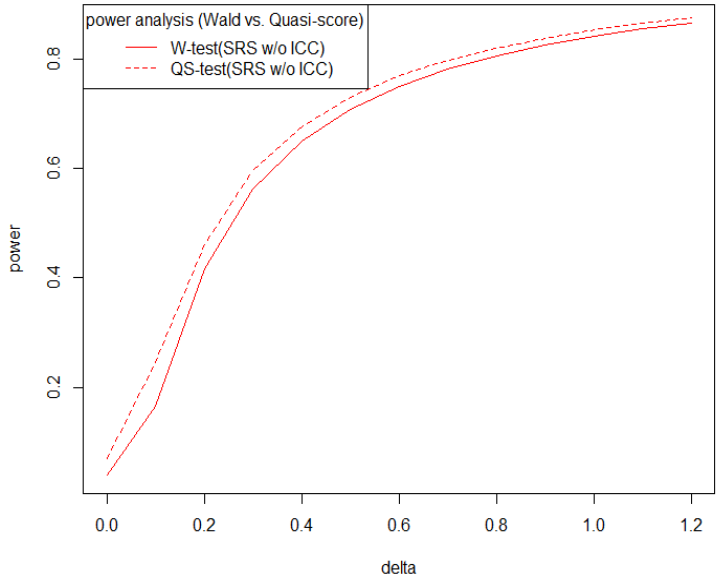


Figure 14: Simluation: Power Analysis (1.2). Wald test and Quasi-score test

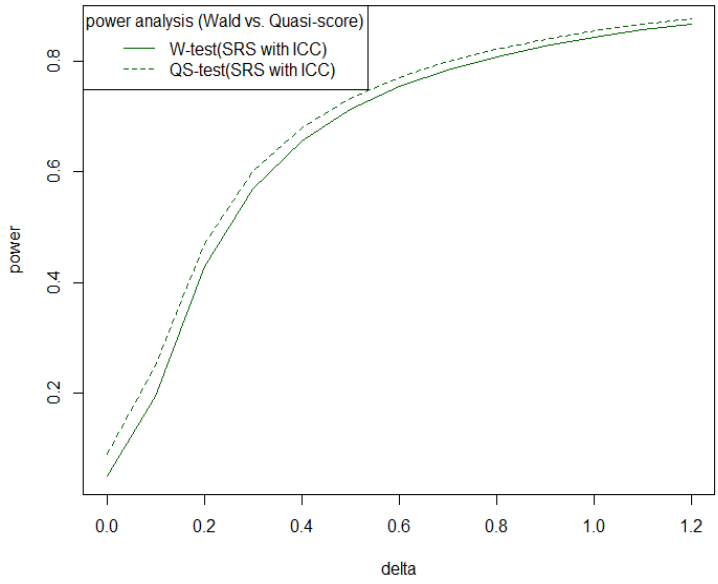
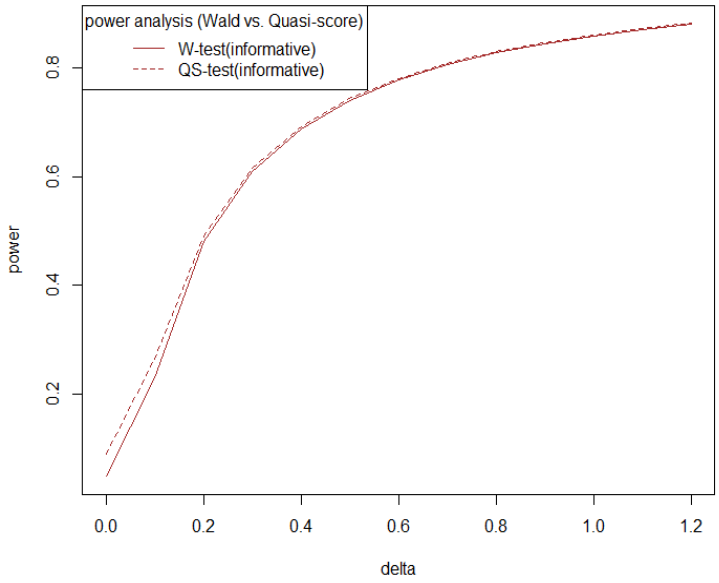


Figure 15: Simluation: Power Analysis (1.3). Wald test and Quasi-score test



From Figure 13 and Figure 14, we find that the power calculated by Quasi-score test is always higher than the one calculated by Wald-test under SRS with $ICC = 0$

and SRS with $ICC \neq 0$ apprroaches, especially when Δ is small. And, under informative sampling, power calculated by Quasi-score test is slightly higher than the one calculated by Wald-test when Δ is smaller than 0.2, and then there is almost no power difference as Δ increases based on Figure 15. Additionally, Type I error is more conservative if using Wald-test than using Quasi-score test, given by Table 32 and all the figures.

Chapter 7 Summary

In this chapter, we summarize the contributions we made based on our extended QEM methods and the important findings from our simulations and analysis of the 1976 NHIS diabetes data. Research limitations and areas of future will be presented at the end of this chapter.

QEM has been used for estimating familial aggregation as recurrence risk when all the families in a sample are the same size. We apply more recent research that used marginal approach for recurrence risk studies when family sizes are varied but only for sibling relationship. We identified this estimation as being based on composite likelihood. Because the recurrence risk considers two family members disease status, we based our estimation of recurrence risk on a pairwise composite likelihood.

Our research made extensions of original QEM and built two new models: race-EQEM and relationship-EQEM. For these two models, we added family relationship and race to the original QEM respectively, and then the new models can be applied for different family relationships or different races. Family relationship is an example of pairwise-level covariate and race is an example of family-level covariate. Thus, studies about pairwise-level covariate effects on recurrence risk can be solved using relationship-EQEM, and similary studies about family-level covariate effects on recurrence risk can be solved using race-EQEM. As for the individual-level covariate, we can treat them as categorial variables and use relationship-EQEM to estimate the model parameters, or we can use propensity score weighting to adjust for the individual-level confounders. We assumed there were only Father-Child, Mother-Child and Child-Child relationships and only whites and other race for family pairs for simplification, however, our models can be applied for any family relationships and races.

Our research successfully extented the original QEM by including simultaneous estimation of recurrence risk for multiple relationships (called relationship-EQEM). We also allowed the parameters of the model to differ according to family-level categorical covariates such race (called race-EQEM), assuming the race was approximately the same across family members. This permitted simultaneous estimation of recurrence risk between family relatives by different family-level covariates. This led to conduct hypothesis tests among model parameters and function model parameters such as recurrence risk.

Another direction that we extended the QEM was accounting for individual-level covariates such as age or gender. One approach for doing this was to treat these covariates as categorical variables and to incorporate them as different parameters in same way we treated different family relationships. This has the unfortunate result of increasing the number of parameters in a model. To avoid this we used propensity score weighting to adjust for the individual-level variables, which can confound the model parameter or recurrence risk estimation. Logistic regression was used to derive the propensity score weights in context of the ATE method. We found methods for adjusting for individual-level covariates to work well in simulations. It is worth noting that even though we considered only Father-Child, Mother-Child and Child-Child relationships and only whites and other race as family -level covariates that our models can be applied for any family relationships and any categorical family-level covariates.

The EQEM can be applied to simple random samples. We also extended the parameter estimation and variance estimation along with hypothesis testing of model parameters and recurrence risk to apply to data from complex samples such as stratified multistage cluster sampling used in household surveys like the NHIS. One way

to collect disease information in household surveys is to use network sampling. This was done in the 1976 NHIS survey to collect diabetes status for the sampled probands in the survey as well as certain first-degree relatives (siblings and parents). Network sampling differs from conventional sampling in that both the sample weight (of the proband) and network weight need be incorporated in the model parameter and variance estimation.

We explored both Wald tests and Quasi-score tests with complex survey data to test hypotheses about model parameters. Both of them can be derived from weighted pseudo (composite) likelihoods and take account of complex survey design features involving clustering and unequal sample weightings. An F-distribution was used as a referent distribution for both Wald tests and Quasi-score tests since it can perform better than a chi-square in approximating the asymptotic distributions of Wald tests and Quasi-score tests for complex survey data (Thomas and Rao, 1987). Simulation studies were conducted to examine Type I errors and power of these tests. The results showed that the Type I error was maintained at the nominal level and the power was very comparable between the two types of tests.

We successfully applied our EQEM methods to the 1976 NHIS diabetes data to study recurrence risk of diabetes and family relationship and race. We compared results of our models with nonparametric survey methods approach proposed by Graubard and Sirken (2013) and obtained the consistent results when estimating the sibling recurrence risk. Wald tests and Quasi-score tests were applied in 1976 NHIS diabetes data for specific hypothesis testing.

A summary of some of the interesting findings of the analysis of family recurrence risk of diabetes in the 1976 NHIS data is given below (Note that ordering exists for

recurrence risk for parent-child relationships):

- Recurrence risk of diabetes conditional on parent's disease status is much lower than the one conditional on child's disease status (i.e., $\hat{RR}_{FC} < \hat{RR}_{CF}$ and $\hat{RR}_{MC} < \hat{RR}_{CM}$) when only living family members are included or dead/alive family members are included in the relationship-EQEM (See Tables 20,22,24). This means if one individual is diabetic, there is a higher probability that his/her parent is diabetic than that his/her child is diabetic.

- Recurrence risk of diabetes between Father-Child is lower than recurrence risk between Mother-Child no matter whether conditioning on child's disease status or conditioning on parent's disease status (i.e., $\hat{RR}_{FC} < \hat{RR}_{MC}$ and $\hat{RR}_{CF} < \hat{RR}_{CM}$)(See Tables 20,22,24). This result works when only living family members are included or dead/alive family members are included in the relationship-EQEM. This could be results of differential reporting error because children are often closer to their mothers than fathers and then they are more likely to report correct information about mothers' diabetes status rather than fathers' diabetes status.

- There is almost the same for $\hat{RR}_W$ and $\hat{RR}_O$ when only including living siblings in the race-EQEM (See Table 27). While recurrence risk of diabetes for whites becomes higher than other races if including both alive and dead siblings in the race-EQEM (i.e., $\hat{RR}_W > \hat{RR}_O$)(See Table 28). Comparing $\hat{RR}_W$ and $\hat{RR}_O$ under only living siblings and both alive and dead siblings, sibling recurrence risk for other races becomes smaller, while sibling recurrence risk for whites almost stays the same after including dead siblings in the race-EQEM. This indicates that the recurrence risk is much lower among dead siblings for other races. It is

possible that people have less chances to remember diabetes status about dead siblings correctly than living siblings, and this reporting error might be larger within other race groups.

While conducting the research for this thesis we identified several areas that needed further work. One was how to model continuous individual-level covariates in the EQEM such as age of an individual. For example when studying including age in the EQEM. As age is usually an important factor for incidence of a disease we would expect it to be related to familial aggregation. In this thesis, we have focused on inference of recurrence risk of phenotypes within families where the phenotype is a binary variable, in particular, prevalence of disease. Because in our extended models we need use pairs of individuals, it is difficult to model age as a continuous variable at the individual-level. Our solutions of accounting for age through propensity score weighting or converting age to a categorical variable and include parameters in the EQEM for combinations of age categories does not treat model age as continuous at the individual-level. In our diabetes example, we adjust for age as a continuous measured confounder when comparing recurrence risk between racial groups using propensity weighting. Another possible approach to account for age would be to estimate recurrence risk in a survival analysis setting where a cohort of families is longitudinally followed for incidence of the phenotype as each individual in the cohort ages. How to model recurrence risk in such a survival analysis setting for data from a longitudinal cohort or complex survey sample is an area of further research.

Another area of future research is to develop Quasi-score tests for directly testing hypotheses about recurrence risks. In this thesis, we provided Wald tests for directly testing recurrence risks. However, there are limitations of Wald tests, e.g., they are not invariant to nonlinear transformations of the parameters and can have poor prop-

erties for small sample sizes.

In addition, in this thesis, we estimated recurrence risk on a population level without considering the effect of family size. More works are needed to see whether family size has influence on recurrence risk.

References

[1] Austin PC. *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies*. Multivariate Behavioral Research, 2011; 46(3): 399424.

[2] Ayala FJ, Fitch WM. *Genetics and the Origin of Species: An Introduction*. Proceedings of National Academy Science, 1997; 94(15): 76917697.

[3] Baker SG. *A sensitivity analysis for nonrandomly missing categoricala dta arising from a national health disability survey*. Biostatistics, 2003; 4(1): 41-56.

[4] Bensen JT, Liese AD, Rushing JT, Province M, Folsom AR, Rich SS, Higgins M. *Accuracy of proband reported family history: the NHLBI Family Heart Study*. Genet Epidemiol, 1999; 17: 141-150.

[5] Binder DA, Patak Z. *Use of estimating functions for estimation from complex surveys*. Journal of the American Statistical Association, 1994; 89(425): 1035-1043.

[6] Boos DD. *On generalized score test*. The American Statistician, 1992; 46(4): 327-333.

[7] Butcher JN, Mineka S, Hooley JM. *Abnormal Psychology, 16E*. Pearson Education, Inc., ISBN 9780205944286, 2014.

[8] Casella G, Berger RL. *Statistical Inference, 2E*. Pacific Grove, CA : Thomson Learning, ISBN 9780534243128, 2002.

[9] Cox DR, Reid N. *A note on pseudolikelihood constructed from marginal densities*. Biometrika, 2004; 91(3): 729-737.

[10] Demnati A, Rao JNK. *Linearization variance estimators for model parameters from complex survey data*. Survey Methodology, 2010; 36(2).

[11] DuGoff EH, Schuler M, Stuart EA. *Generalizing observational study results: applying propensity score methods to complex surveys*. Health Services Research, 2014; 49: 284303.

[12] Granovetter M. *Network sampling: Some first steps*. American Journal of Sociology, 1976; 81(6): 1287-1303.

[13] Graubard BI, Korn EL. *Modelling the sampling design in the analysis of health surveys*. Statistical Methods in Medical Research, 1996; 5: 263-281.

[14] Graubard BI, Korn EL. *Conditional Logistic Regression With Survey Data*. Statistics in Biopharmaceutical Research, 2011; 3(2).

[15] Graubard BI, Sirken MG. *Estimating Sibling Recurrence Risk in Population Sample Surveys*. Human Heredity, 2013; 76(1):18-27.

[16] Guo SW. *Inflation of Sibling Recurrence-Risk Ratio, Due to Ascertainment Bias and/or Overreporting* The American Journal of Human Genetics, 1998; 63(1).

[17] He W, Yi GY. *A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models*. Statistica Sinica 21 (2011), 207-229.

[18] Hunsberger S, Graubard BI, Korn EL. *Testing logistic regression coefficients with clustered data and few positive outcomes*. Statistics in Medicine, 2008; 27(8): 1305-1324.

[19] Imbens GW. *Nonparametric estimation of average treatment effects under exogeneity: A review*. The Review of Economics and Statistics, 2004; 86(1): 429.

[20] Kish L. *Survey Sampling*. New York: John Wiley and Sons, Inc., ISBN 0-471-10949-5, 1965.

[21] Kolenikov S. *Resampling variance estimation for complex survey data*. The Stata Journal, 2010, 10(2): 165-199.

[22] Korn EL, Graubard BI. *Analysis of Health Surveys*. New York: John Wiley and Sons, Inc., ISBN 0471137731, 1999.

[23] Korn EL, Graubard BI. *Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t statistics*. The American Statistician, 1990; 44(4): 270-276.

[24] Kotz S, Johnson NL, Read CB. *Encyclopedia of statistical sciences*. New York: John Wiley and Sons, Inc., ISBN 0471150444, 1982.

[25] Larribe F, Fearnhead P. *On composite likelihood in statistical genetics*. Statistica Sinica, 2011; 21(1): 43-69.

[26] Li CC. *Population Genetics*. University of Chicago Press, 1955.

[27] Li Y, Graubard BI, DiGaetano R. *Weighting methods for population-based casecontrol studies with complex sampling*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2011; 60(2): 165185.

[28] Liang KY. *Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models*. Biometrics, 1987; 43(2): 289-299.

[29] Liang KY, Zeger SL. Kung-Yee Liang; Scott L. Zeger *Longitudinal data analysis using generalized linear models*. Biometrika, 1986; 73(1): 13-22.

[30] Lumley T. *Analysis of complex survey samples*. Journal of Statistical Software, 2004; 9(1).

[31] Matthews AG, Finkelstein DM, Betensky RA. *Analysis of familial aggregation in the presence of varying family sizes*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 2005; 54(5): 847-862.

[32] Matthews AG, Finkelstein DM, Betensky RA. *Analysis of familial aggregation studies with complex ascertainment schemes*. Statistics in Medicine, 2008; 27(24): 50765092.

[33] Morgan SL, Todd JJ. *A diagnostic routine for the detection of consequential heterogeneity of causal effects*. Sociological Methodology, 2008; 38:231281.

[34] Nelson RB. *An Introduction to Copulas*. Springer, 2006, ISBN 978-0-387-28678-5.

[35] Olson JM, Cordell HJ. *Ascertainment bias in the estimation of sibling genetic risk parameters*. Genetic Epidemiology, 2000; 18(3).

[36] Prentice RL. *Correlated binary regression with covariates specific to each binary observations* Biometrics, 1988; 44(4): 1033-1048.

[37] Rao JNK, Scott AJ, Skinner CJ. *Quasi-score tests with survey data*. Statistica Sinica, 1998; 8(4): 1059-1070.

[38] Ridgeway G, Kovalchik SA, Griffin BA, Kabeto MU. *Propensity Score Analysis with Survey Weighted Data*. Journal of Causal Inference, 2015; 3(2): 237-249.

[39] Risch N. *Linkage Strategies for Genetically Complex Traits: 1. Multilocus Models*. American Journal of Hunman Genetics, 1990; 46(2).

[40] Rosenbaum PR, Rubin DB. *The central role of the propensity score in observational studies for causal effects*. Biometrika. 1983; 70(1): 4155.

[41] Sirken MG. *Encyclopedia of Biostatistics—Network Sampling*. John Wiley and Sons, Ltd, 1998; 2977.

[42] Sirken MG, Graubard BI, McDaniel MJ. *National network surveys of diabetes*. Proceedings of the American Statistical Association of the Section on Survey Research Methods, American Statistical Association, 1978, 631-635.

[43] Sklar A. *Fonctions de rpartition n dimensions et leurs marges*. Publications de l'Institut de statistique de l'Universite de Paris, 1959, 8: 229231.

[44] Thomas DR, Rao JNK. *Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling*. Journal of the American Statistical Association, 1987; 82(398): 630-636.

[45] Varin C, Reid N, Firth D. *An overview of composite likelihood methods*. Statistica Sinica, 2011; 21(1): 5-42.

[46] Zanutto EL. *A comparison of propensity score and linear regression analysis of complex survey data*. Journal of Data Science, 2006; 4: 6791.

[47] Zou G, Zhao H. *The estimation of sibling genetic risk parameters revisited*. Genetic Epidemiology, 2004; 26(4).