A Method for Pedigree-based Propagated Risk Phenotyping on Clalit Data

Nancy-Sarah Yacovzada, Iris Kalka, Eran Segal, Eran Hornstein
Weizmann Institute of Science, Israel

Motivation

Family history is known to be one of the most important disease risk factors necessary for the implementation of precision medicine (Aronson and Rehm, 2015; Guttmacher et al., 2004). While previous research has focused on family studies of known relatives, primarily twins, Electronic health records (EHRs) offer a new alternative to traditional phenotyping and an opportunity for observational epidemiological studies, allowing to investigate not only twins, but the entire families across several generations. Clalit data has the unique benefit where relationships between parents and child are explicitly documented - enabling for construction of family-trees and rapid risk assessments at scales that were previously impossible to achieve. EHR-based risk estimates are particularly well suited for complex traits that require large numbers of patients. Moreover, our approach allows for hereditary risk evaluation of diseases not previously investigated in family-based or twin studies. Importantly, our work is based on family-trees representative of different Israeli sub-populations, while most studies have predominantly involved white Europeans.

Pedigree-based Hereditary Risk Score

Here, we propose a novel approach to estimate one's propensity to a given phenotype Y based on Clalit data, and demonstrate a utilization of this estimate as the "phenotypic score" (outcome) for Parkinson.

In standard EHR-phenotyping approach in case-control cohorts, "cases" are usually defined by a clear diagnosis with the relevant ICD code (i.e., $Y_i \in \{0,1\}$). To represent the role that genetics plays in traits across different populations, we developed a risk scoring technique that also incorporates pedigree information, and thereby extends standard phenotyping. Such scoring allows producing a **risk score for offspring of diseased individuals based on known medical records of their ancestors.** The motivation behind the score is as follows: relatives of patients will obtain a higher risk score than relatives of non-diseased ones. Based on diagnosis information, combined with pedigree information, date of birth, sex, and parental clinical information, a **population-scale family trees with millions of relatives were constructed**. The genealogical relationships between individuals is structured into graph topologies and stored as a sparse kinship matrix, allowing to approximate the relatedness of individuals without the need for expensive genetic tests.

We hereby define a propagated **Hereditary Risk Score σ_Y(HRS)**, a continuous score where $\sigma_Y(i) \in [0,1]$, for approximating the predisposition of an individual i to the disease Y according to his/her family context. **The HRS scores can be estimated without the use of**

expensive sequencing data and can be generated for every phenotype $Y$with a significant heritable component.

We hereby constructed approximately hundreds family-trees from Clalit data and computed the kinship coefficients for each of the 11M pairs of relatives (Kalka et al, 2020). Then, we marked nodes of known patients ($\boldsymbol{Y = 1}$), propagated risk to their relatives and finally estimated $\sigma_{\boldsymbol{Y}}$. The computation of the kinship coefficients and formulation of proposed scaled risk score are detailed in Methods.

Our proposed approach introduces significant benefits over using a binary outcome variable (disease/healthy, $Y_i \in \{0,1\}$). The use of the continuous σ scores as a proxy of the phenotype increases computation power in ~ one order of magnitude by using considerably larger cohorts (for example, instead of 50K Parkinson cases, a continuous PD-risk score is established for >500K subjects). In addition to the information gained by avoiding dichotomized information, using a scaled continuous risk score presents an ordinal over binary outcome variable, allowing ranking and stratification into risk percentiles. Moreover, **our technique introduces a whole new approach for epidemiological questions with regard to early-life risk factors in offspring of patients**. The HRS score can be used to define groups with higher risk and to identify events present before an individual reaches the symptomatic stage and is diagnosed.

Methods

Cohort definition: Clalit EHR Data

We used EHRs of Clalit Health Services (Clalit), Israel's largest healthcare provider, including more than five million people (over 50% of Israel's population) with 18 years of longitudinal measurements dating back to 2002. Patients' data includes full clinical information including diagnoses, lab tests, prescribed medication, basic demographics and data from over 1,500 clinics and 14 hospitals (30% of Israeli hospital beds).

Pedigree and Kinship Matrix

To model the covariance between individuals in Clalit's data, in the absence of DNA sequencing data, we organized data into graph topologies that preserve the genealogical relationships between individuals and extracted population-scale family trees. The graphs' corresponding kinship matrix Φ is calculated directly from the data, where $\phi(i,j)$ is the genetic resemblance between individuals i and j. In more detail, the **coefficient of kinship** ϕ, is a probabilistic estimate that a randomly selected allele from two individuals i and j at a given locus will be identical by descent (IBD), assuming all founder alleles are independent. For N subjects, these probabilities can be assembled in a $N \times N$symmetric matrix termed the **kinship matrix** Φ (also referred as "*Additive Relationship Matrix*"), with the coefficient of kinship $\phi(i,j)$ as elements. The ϕ coefficients are estimated according to a unique shortest path between them through a shared ancestor. Such each path increases the similarity between the pair of individuals by $2^{-L(p)}$, where $L(p)$ is the number of edges in the path. Thus, $\phi(i,j)$ is a measure of the degree of biological relationship between two individuals i and j, and it is obtained by a summation of number of edges by which i and j are connected to their common ancestors. The $\phi(i,j)$ coefficients are then stored as a sparse kinship matrix, allowing to perform analytical procedures efficiently despite the large scale of Clalit data (Shor et al. 2019).

Formally, let $R(i)$ be the group of adult relatives of subject i. If $j \in R(i)$, then subject j shares a common ancestor with subject i, and the **cumulative relationship** of subject i in the family tree is represented by:

$$\sum_{\forall j \in R(i)} \phi(i,j)$$

If $j \notin R(i)$, then we define $\Phi_{(i,j)} = \phi(i,j) = 0$. The $\phi(i,j)$ coefficient can be simplified and approximated by,

$$\Phi_{(i,j)} = \phi(i,j) = \sum_{p} 2^{-L(p)}$$

giving p enumerates all paths connecting i and j with unique common ancestors and $L(p)$ is the length of the path p.

Hereditary Risk Estimates as Outcome from Kinship Matrix

Denote $\sigma_{\boldsymbol{Y}}(\boldsymbol{i})$ **Hereditary Risk Score (HRS)**, the conditional risk of disease $\boldsymbol{Y}$ to appear in an offspring i, given his ancestors' medical background. The conditional risk is then normalized with the cumulative familial relationships in the population, and provides an estimate of the excess hereditary risk of a disease Y.

For that purpose, we distinguish between two types of nodes in the family tree:

1. Case-nodes ($Y(i) = 1$, e.g., subject i diagnosed with Parkinson), and
2. Control-nodes ($Y(i) = 0$, subject i was not diagnosed with Parkinson).

If subject i was diagnosed with disease $\boldsymbol{Y}$ (meaning $Y(i) = 1$), then $\sigma_{\boldsymbol{Y}}(\boldsymbol{i}) = \mathbf{1}$.
If $Y(i) = 0$, and if subject i has a diseased ancestor j (i.e., $Y(j) = 1$), then we expect $\sigma(i)$ to increase by the coefficient of relationship $\phi(i,j)$. Thus, total absolute risk of subject i for phenotype Y is represented by summation of $\phi(i,j)$ for each $j \in R(i)$ and $Y(j) = 1$. We then normalize the total risk by the cumulative relationships of i with all nodes in $R(i)$.

Formally, the **hereditary risk score** of subject i for phenotype Y is obtained by:

$$\sigma(i) = \frac{\sum_{j \in R(i), Y(j)=1} \phi(i,j)}{\sum_{j \in R(i)} \phi(i,j)}$$

Potential Early-Life Risk Factors and Hereditary Risk Score of Parkinson Disease among Clalit members

Introduction

Neurodegenerative diseases, such as Parkinson's disease (PD), Alzheimer's disease (AD), and Amyotrophic Lateral Sclerosis (ALS) are major public health issues in the aging population. Although genetic analysis of these diseases has also been vastly described, yet the mechanisms governing the extent of vulnerability to such diseases remains unresolved. Findings of research obtained over the past 20 years suggest that neurodegeneration diseases may have its origins in early life. The central nervous system undergoes considerable degeneration for years before the onset of neurodegenerative symptoms, such as Mild Cognitive Impairment (MCI) in AD and tremors in PD.

PD has been considered a sporadic disease for a long period, and only 15% of patients have family history (Farrer et al., 2006). However, multiple family aggregation studies supported that the relatives of PD, especially the siblings of patients, had a higher risk of PD than the relatives of non-PD patients [Thacker et al., 2008]. Also, the first-degree relatives of patients with PD are generally at the increased risk of nonmotor symptoms, such as anxiety, depression, and dementia. Pathological studies have shown a 40%–60% threshold of dopamine neuron loss in the SN pars compacta (SNc) and 60%–70% striatal dopaminergic reduction before the appearance of typical motor symptoms meeting PD diagnostic criteria [Fearnley et al., 1991]. An International Parkinson Disease and Movement Disorder Society (MDS) [Stern et al., 2012] task force proposed the terminology for the early stage of PD, that is, prodromal PD, in which affected subjects might have some NMS and/or subtle motor signs, but no typical motor symptoms meeting the diagnostic criteria for PD are observed [Siderowf et al., 2012 ; Rožanković, 2020]. Early symptoms of Parkinson's may be experienced months or even years before the traditional motor symptoms emerge, such as constipation and digestion problems, reduced sensitivity to odors (hyposmia), loss of smell (anosmia), idiopathic rapid eye movement behavior disorder (RBD), etc. Here, we wish to evaluate several environmental factors, blood lipids, clinical data and biomarkers capable of revealing the development of the disease in advance to loss of dopaminergic neurons. The non-motor clinical features and markers could be used to screen for PD before motor symptoms are present, and to identify PD before it reaches symptomatic stage [Biomarkers of Parkinson's disease: Present and future].

Hereditary Risk Score Estimates of Parkinson's Offspring among Clalit members

We hypothesize that **early-life risk factors of the disease can be discovered by detecting pre-symptomatic events and early markers that are more prevalent in offspring with diseased ancestors than in non-diseases ones**. Such events and markers can introduce novel prediction tools for early diagnosis in the pre-symptomatic phase of the disease.
To explore such hypotheses, we suggest first generating Hereditary Risk Score for each member in Clalit, according to his propagated PD-risk from his ancestors. Therefore, we **utilize here our proposed pedigree-based hereditary risk approximation** to estimate the continuous risk score $\sigma(i)$ as a proxy of the risk of the individual i to develop PD. This score is used for estimating the **observed and causal effect** of different exposures, markers, events and clinical information on σ_Y (instead on the binary Y).