Risk prediction and Type II Diabetes

Nicholas A. Furlotte[1], Sheel Dandekar[1], Robin Smith[1], Nicholas Eriksson[1] and David A. Hinds[1]

1. 23andMe, Inc., Mountain View, CA, USA

Abstract

As the prevalence of Type II Diabetes (T2D) continues to increase worldwide, the ability to accurately assess the risk for developing this disease is becoming increasingly important in clinical practice. Accounts vary as to the clinical utility of genetic risk prediction models for T2D. Estimates of heritability vary widely, but it is generally accepted that environmental factors such as food consumption, exercise frequency and body mass index play a much larger role in the development of this disease than the genetic variations implicated through genome-wide association studies. Furthermore, the effects of genetic variations may only be apparent under particular environmental conditions – so called gene-by-environment interactions. We explore these issues in a cohort of over 300,000 23andMe customers. We evaluate the predictive power of the most recent and robust genetic variations implicated in the etiology of T2D through association analysis and compare their predictive ability with environmental predictors related to individual behavior. In addition, we search for genetic variations exhibiting environmentally specific genetic effects and quantify the proportion of the total trait variation attributed to these gene-by-environment interactions.

Overview

- All participants were drawn from the customer base of 23andMe, Inc., a consumer personal genetics company. Customers were genotyped on the Illumina HumanOmniExpress+ platform.
- Phenotyping was accomplished through online surveys. For example, customers are asked if they have ever been diagnosed with type 2 diabetes.
- Using the genetic and phenotypic information, we conducted a genome-wide association study and combined the results of this analysis with previously published studies in order to obtain a set of SNPs to be used as genetic predictors of disease.
- Using age, sex, BMI and principal components as baseline or non-genetic predictors, we evaluate the gain in predictive accuracy achieved when adding genetic predictors.
- Next, we evaluate the calibration of the genetic model both in the training set and in an independent test set (in sample vs out of sample prediction).
- Finally, we examine the potential role of gene-by-environment (GxE) interactions.

Discussion

In this work, we evaluated the predictive accuracy of a standard logistic regression based model. We found that the genetic effects are small and do not provide a huge increase in predictive power when compared with demographic predictors such as age, sex and BMI. However, we also show that genetic risk may be used to differentiate individuals within the same demographic risk group. This potential for differentiation could help guide individuals and clinicians in their health decisions. In addition, we show that although many environmental and lifestyle traits are associated with T2D risk, the magnitude of genetically-dependent environmental effects remains unclear.

Acknowledgments

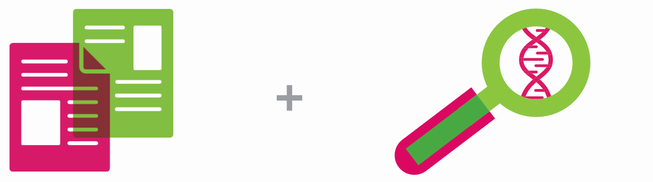
We thank 23andMe customers who consented to participate in research for enabling this study. We also thank employees of 23andMe who contributed to the development of the infrastructure that made this research possible.

Methodology

Study Cohort

Our study cohort consists of about 165,000 23andMe customers aged 45 years of age or older, of European descent and who have indicated their T2D status (diagnosed or undiagnosed). Each individual has also reported their current height and weight from which we derive BMI.

	Case	Control
Male	7,233	78,683
Female	4,701	74,291



SNP Selection

We started with a set of 58 known T2D associated SNPs [1] and augmented this set with 22 independent loci identified via a 23andMe GWAS. GWAS SNPs were selected if they had a p-value <= 5e-8 and were more than 500kb from a previously known association.

Discovery GWAS

The total cohort was split into a training set consisting of 80% of the original cohort and a test set consisting of 20% of the original cohort. Related individuals were removed so that each pair of individuals are unrelated within and between data sets.

Predictive Model

In order to define a predictive model for T2D, we fit a basic logistic regression. The log of the odds of reporting T2D is assumed to be a linear function of age, sex, BMI and the first five population principal components (all represented in matrix notation as X below).

$$log(\frac{p(T2D|X)}{1-p(T2D|X)}) = X\beta + \epsilon$$

Evaluating the Model

Discrimination

Evaluated using ROC/AUC, discrimination measures how well the model separates those who have reported being diagnosed with T2D and those who did not.

Calibration

Calibration measures how accurate estimated probabilities are by comparing the expected and observed probability of disease in risk deciles. Expected probability is calculated as the mean estimated probability within a decile. This is compared to the observed case frequency within the same decile.

Discrimination

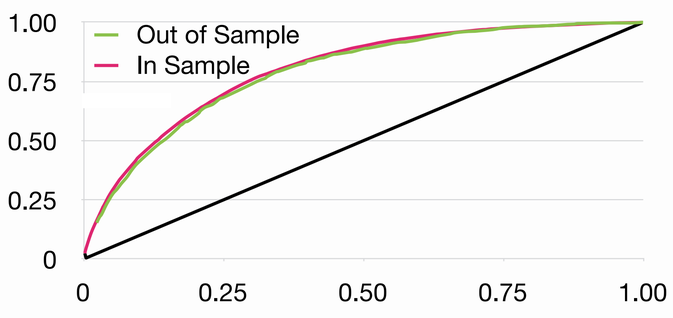


Figure 1. In Sample and out of sample prediction result in similar ROC curves. We find that the ROC curves are very similar for in sample and out of sample. The AUCs are 0.80 and 0.79, respectively.

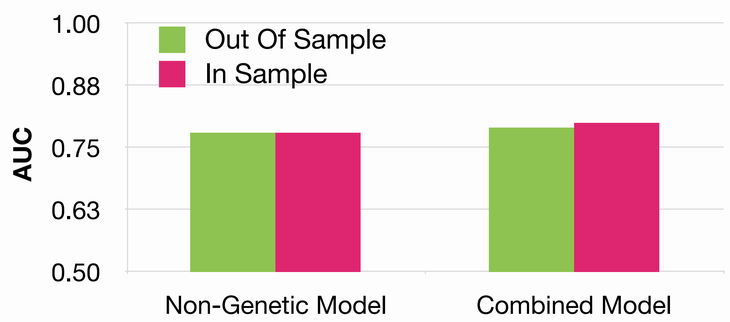


Figure 2. Genetic predictors add little to the discriminatory power when compared with age, sex and BMI.

Model Calibration

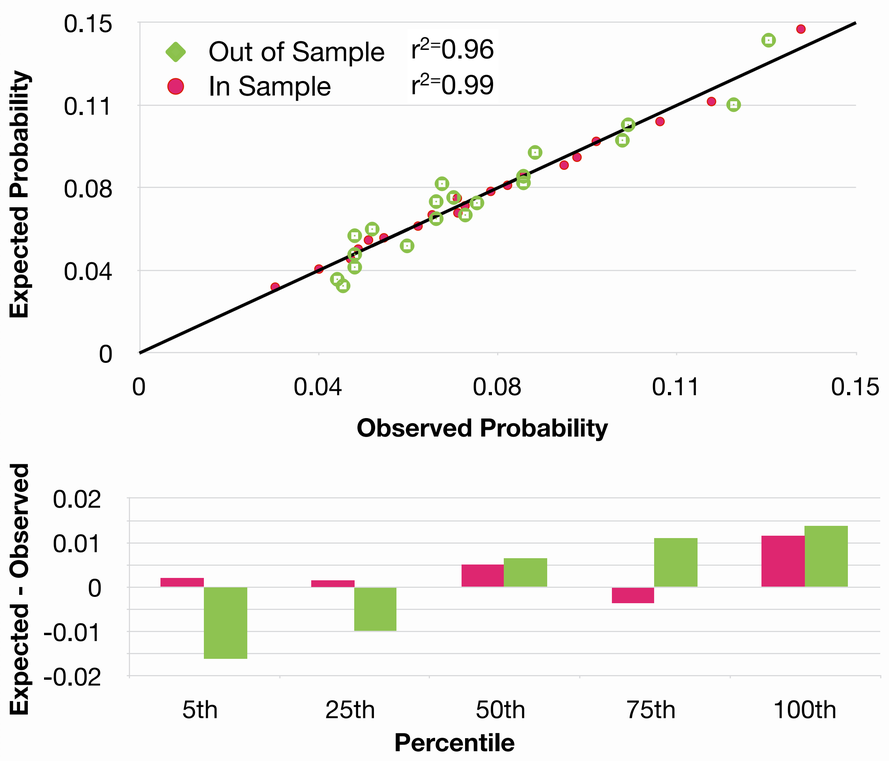


Figure 5. The model is well calibrated and only slightly overfit.
Calibration is evaluated by plotting the average predicted probability in each group of individuals by the observed frequency of disease in each group. Groups are determined by splitting individuals into 20 equally sized bins after ranking them by their **genetic risk**. The in sample values track closely to the diagonal line, indicating that this model is well calibrated (expected is close to observed). Not surprisingly, the out of sample values do not track this line as closely, indicating that out of sample prediction is less calibrated and that the model is overfit. The bottom plot shows a different view of this by plotting the deviations from the diagonal line.

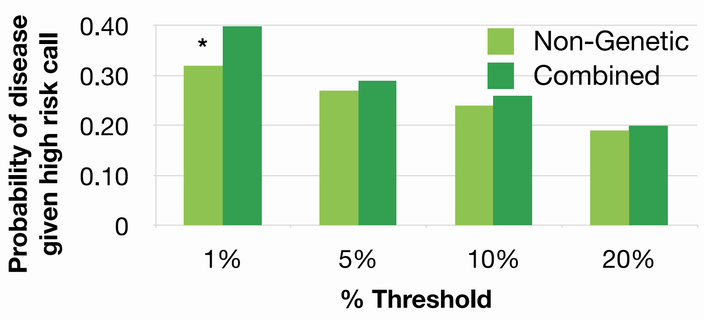


Figure 4. A simple genetic risk based classifier could result in clinically relevant information. Clinical relevance is difficult to determine. Here we use a simple classifier that calls people in the top x% of the risk distribution as high risk and those below as typical risk. The probability of disease given a high risk call is then compared between the two models.

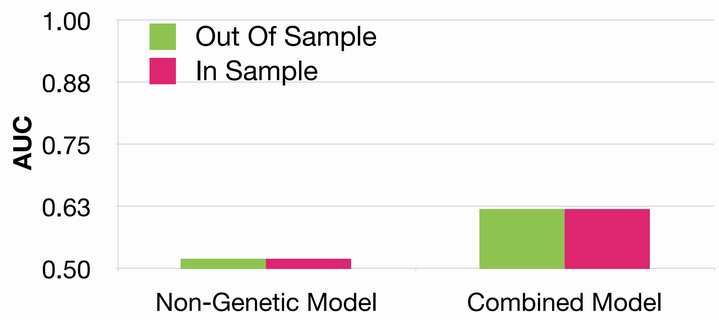


Figure 3. The predictive power of genetics is seen more clearly by computing AUC in risk strata. Here we compute the average AUC for the non-genetic and combined models across risk strata. Each risk strata is defined by sex and BMI and age range.

Gene By Environment

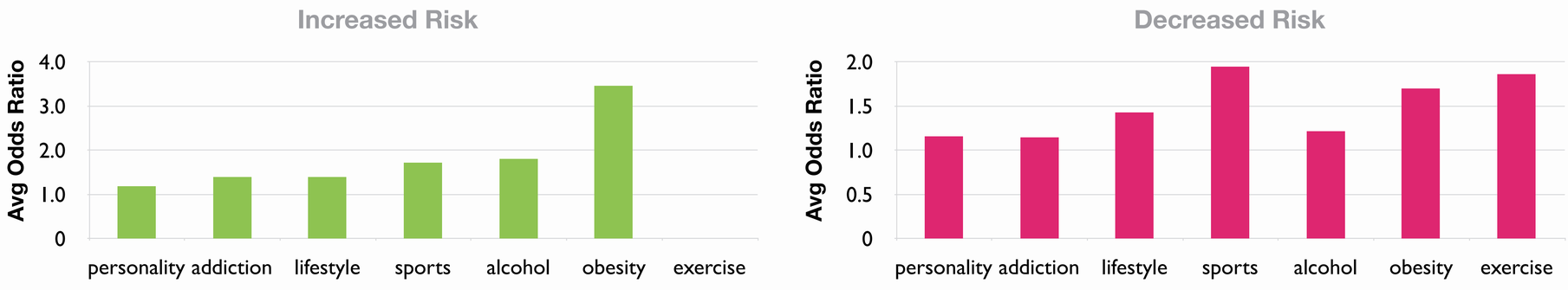


Figure 6. Many factors are strongly associated with T2D risk. We found roughly 400 phenotypes that were significantly associated with T2D. Here we report the odds ratios resulting from averaging effects across phenotypes within categories. The left plot was generated with phenotypes that increased risk and the right with phenotypes that decrease risk.

Gene-by-environment (GxE) GWAS

We performed a simple GxE analysis by looking for interaction effects between SNPs and environmental conditions. For SNPs, we used the set of of genetic predictors used throughout. For environmental conditions, we chose the set of phenotypes that were significantly associated with T2D. Each pair was tested for association with T2D by fitting an interaction term in the standard logistic regression.

Although, we found a handful of significant associations, the effects are very small. As a result, the extent to which T2D risk is influenced by specific GxE interactions remains unclear.