A survey of integrating age-at-onset genetics for predicting the age-specific disease risk: Polygenic Hazard Score

Chao Tian*, the 23andMe Research Team, David Hinds
23andMe, Inc., Mountain View, CA research.23andme.com * ctian@23andme.com

Introduction

Human disease is characterized by marked genetic heterogeneity, which has important implications for gene discovery. In particular, evidence suggests distinct genetic susceptibility between early-onset and late-onset diseases. In this study, we demonstrated that an association study of age-at-onset information collected in the case-only cohort could lead to the discovery of novel genetic risk factors that were not detectable in case-control GWAS (GWAS_cc), despite reduced sample size. Table 1 shows a few examples. A missense variant (rs28929474) in SERPINA1 is highly significant in GWAS_age. Further investigation found that rs28929474 is associated with only early-onset high cholesterol and does not significantly contribute to late-onset HC. The LDscore Genetic correlation between GWAS_cc and GWAS_age is 0.76 (pvalue = 1.85e-155, sdp=0.00).

Results

Genome Wide Association study on age-at-onset in case-only cohort (GWAS_age)

Age-at-onset information collected in the HC case-only cohort leads to the discovery of a few novel genetic risk factors that were not detectable in case-control GWAS (GWAS_cc), despite reduced sample size. Table 1 shows a few examples. A missense variant (rs28929474) in SERPINA1 is highly significant in GWAS_age. Further investigation found that rs28929474 is associated with only early-onset high cholesterol and does not significantly contribute to late-onset HC. The LDscore Genetic correlation between GWAS_cc and GWAS_age is 0.76 (pvalue = 1.85e-155, sdp=0.00).

SNP discovery

1245 independent SNPs were associated with increased risk of HC (GWAS_cc), and 1459 independent SNPs were associated with age-at-onset of HC (GWAS_age), with P<10^-8. The two sets of SNPs were entered into COX proportional hazard model for computing PHS_ccSNP and PHS_ageSNP. It may also be of interest to consider the performance of a traditional polygenic risk score (PRS), built with their corresponding odds ratios (OR) and effect sizes (beta) from the discovery GWASes. We conducted this post-hoc analysis and found that the performance of re-trained coefficients with training dataset performs better. In particular, the training dataset is used more similar to the dataset for predictive analysis.

Cox proportional hazard model

In the independent training dataset, the graphical comparisons among Kaplan-Meier estimates and Cox proportional hazard model stratified on only PHS (combined PHS_ccSNP, PHS_ageSNP) indicate that the proportional hazard assumptions were not severely violated (Figure 1).

Risk prediction with PHS

To verify whether the PHS accurately predicts age at onset of HC, we calculated the PHS for all participants in the test dataset. We calculated a hazard ratio comparing men with high scores (> 90% PHS) with those with low risk (bottom 10% PHS). The time-dependent ROC analysis was done using R 'survivalROC' package using nearest-neighbor estimator.

Discussion

The PHS could incorporate into a personalized assessment of individuals’ age related risk that can guide the decision of whether and when an individual needs to order screening test. Here we carefully investigated the Cox survival model for age-specific risk prediction and provided some practical guidance on incorporating age-at-onset genetics into disease risk prediction. The work still has limitations and further development are required. When using the survival model, one thing to consider is that the baseline hazard estimates derived from GWAS samples may not be used directly. There are a few methods for deriving population-based incidence rate estimates. Previous studies showed that simple inclusion or exclusion of future cases in each risk set induced an under- or over-estimation bias in the regression parameters, respectively. A weighted Cox model was suggested for accounting for age conditional probabilities of developing the disease in the source population may provide less biased estimation. When incorporating multiple correlated PHS into a single model, a regularized regression or using a derived weighted multi-trait PHS may provide more robust results.

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.

References

1. Selmet TR et al. NAM. 2010 Jan 10;360:j5757
2. Tan CH et al. NAM. 2018 Jan 10;360:j5757
3. Seibert TM et al. NAM. 2017 Sep 12;360:j5757-488
4. Lynch T et al. NAM. 2016 Dec 26;360:j5757-488
5. Lutt PJ et al. NAM. 2016 Dec 26;360:j5757-488
7. Henn L et al. NAM. 2016 Dec 26;360:j5757-488
8. Kocher B et al. NAM. 2016 Dec 26;360:j5757-488
9. Meier MM et al. NAM. 2018 Apr 7;360:j5757-488

Copyright © 2018 23andMe, Inc. All rights reserved.