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 factor so-called modifiers of age-at-onset, despite reduced sample size. For example, variants in *SERPINA1, HDAC5* and *CTNND2* were only associated with early-onset high cholesterol (HC) and were detectable in the linear regression analysis on age-at-onset information, but were missed by traditional case-control studies that ignore age-at-onset differences.

This degree of heterogeneity of age-at-onset also has important implications for development of personalized genetic assessment. As a result, we should include modifiers of age-at-onset into the predictive modeling to increase predicting accuracy, especially for earlier onset diseases. Currently, polygenic risk scores (PRS), derived mainly from case-control GWAS, have been used for predictive modeling of many complex disease risks. This approach is considered clinically suboptimal for assessing an age-dependent disease where a subset of "controls" will become "cases" over time. Scientists are trying to extend the work by integrating genetics within a survival-analysis-epidemiology framework to derive a polygenic hazard score (PHS) and hope to use it to inform both whether and when to order screening tests. Some preliminary work has been done for Alzheimer's disease[1] and prostate cancer[2].

Methods

Samples:

Three independent cohorts (**Table 1**) on high cholesterol were used and composed of customers of European descent, who were consented for research and genotyped on the Illumina BeadChip as part of the 23andMe Personal Genome Service. Genotype imputation was done using internally-developed tool, Finch (based on Beagle graph-based algorithm[3]) and also a new phasing algorithm, Eagle2[4]. A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[5].

Name	Case#	Control#	Female %	(0,30]	(30,45]	(45,60]	(60,Inf]
Discovery: GWAS_cc	387413	811537	51%	11%	24%	27.5%	37.5%
Discovery: GWAS_age	274973	NA	48%	1%	10%	28%	61%
Training	114607	394556	55%	18%	25%	26%	31%
Test	22414	77586	54%	18%	26%	26%	31%

Table 1: Cohort table

GWAS

To select HC associated SNPs, we carried out two GWAS: logistic regression GWAS on carrier status of HC (GWAS_cc) and linear regression GWAS on age-at-onset of HC in case-only discovery cohort (GWAS_age).

Polygenic hazard score

The parameters were re-estimated for each associated SNP using COX proportional hazard model (using R 'survival' package) controlling for sex, age and principal components (PCs) in the training dataset and then combined with individual genotypes to generate the two sets of PHS: PHS_ccSNP and PHS_ageSNP (see equation below, where X is the participant's genotype for n selected SNPs and the beta is the corresponding parameter estimates from Cox model). The current age of controls was treated as censored time and the censoring process did not depend on genetics. The final model (model_final) was built in the training dataset with the combined PHS (weighted sum of PHS_ccSNP and PHS_ageSNP with weights being their estimated coefficients), co-varied for the effects of age, sex, and PCs. For comparison, we built the model using only PHS_ccSNP in the similar way (model_ccSNP).

$$PHS_x = \sum_i^n X_i \beta_i$$

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[6].

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 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, sd=0.03).

SNP	cytoband	gene.context	allele	freq	p-value_GWAS_age	p-value_GWAS_cc
rs28929474	14q32.13	[SERPINA1]	C/T	0.018	4.41E-13	1.39E-05
rs433610	17q21.31	[HDAC5]	A/G	0.718	1.10E-09	2.76E-02
rs4578371	11p15.2	[ARNTL]	G/T	0.246	3.16E-08	2.76E-04
rs17801366	5p15.2	[CTNND2]	C/T	0.967	4.63E-08	3.33E-02
rs77960177	4q26	NDST4---[]---MTRNR2L13	D/I	0.61	1.71E-07	5.05E-04

Table 2. Significant SNPs in the GWAS_age and their p-values in GWAS_cc.

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 $P<10^{-3}$. 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 used is 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 only on PHS (combined PHS_ccSNP, PHS_ageSNP) indicate that the proportional hazard assumptions were not severely violated (**Figure 1**).

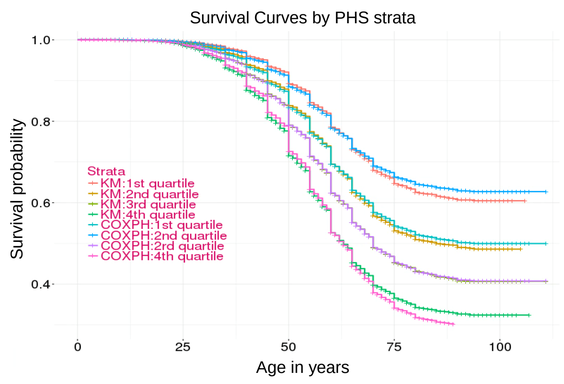

Figure 1. Kaplan-Meier estimates and Cox proportional hazard model with training data stratified by PHS quartiles.

The final cox model was fitted with PHS controlling for age, sex, and PCs, suggesting that PHS was a significant predictor of age-at-onset of HC (z=125.7, HR=1.46, P<1e-16, 95%CI is 1.45-1.47). Being a male increases the hazard by a factor of 1.35 (HR=1.35, P<1e-16). Figure 2 shows the final model fit for males with lower 10% PHS (10%), median (50%) and top 10% (90%) PHS. The plot showed that PHS successfully stratified individuals into different risk strata. For male with the high 90% PHS, at 50% risk, the expected age of developing HC is 65 y; however for male with the low 10% PHS, the expected age of developing HC is approximately 54y. The hazard ratio comparing the 90% PHS to the first 10% is 2.63.

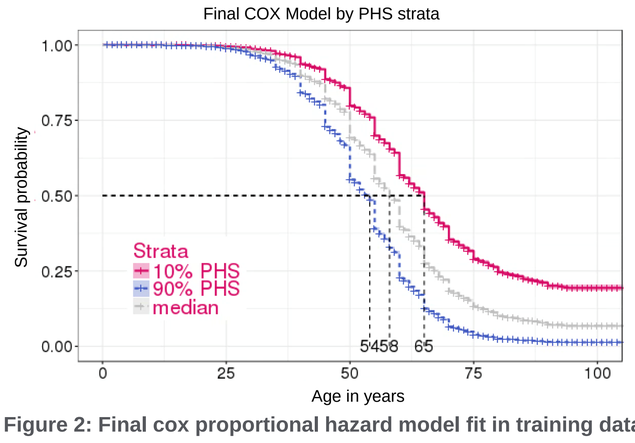

Figure 2: Final cox proportional hazard model fit in training dataset.

Results (cont.)

Model testing

To assess the predictive power, we applied the final model In the independent test dataset. By assuming naively 50% risk for meeting criteria for HC diagnosis, we obtained the predicted age of onset for developing HC, which is correlated with the empirical (actual) age of onset (r=0.6). ROC curves are useful tool for showing diagnostic potential of continuous markers for disease outcomes. PHS could be used as prognostic marker in the setting of survival data. We used the time-dependent ROC curve to evaluate/compare the discrimination potential of PHS for time-dependent disease outcomes. **Figure 3** suggests that the final model with combined PHS (combine PHS_ccSNP and PHS_ageSNP) outperforms the PHS_ccSNP only model.

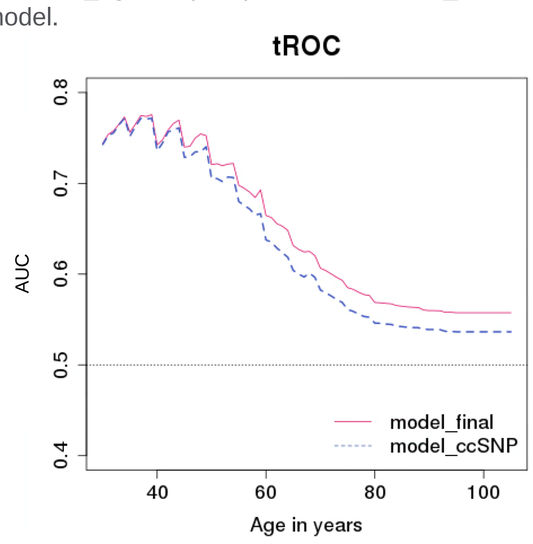

Figure 3. Time dependent Roc analysis.

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 many not be used directly. There are a few methods for deriving population-based incidence rate estimates[7]. 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 that weights subjects according to 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[8] or using a derived weighted multi-trait PHS[9] 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. Seibert TM *et al. BMJ*. 2018 Jan 10;360:j5757
2. Tan CH *et al. Ann Neurol*. 2017 Sep;82(3):484-488.
3. Browning SR *et al. Am. J. Hum. Genet.* **81**: 1084–1097 (2007).
4. Loh PR *et al. Nature Genetics* **48**: 811-6 (2016).
5. Henn BM *et al. PLoS One* **7(4)**: e34267 (2012).
6. Heagerty PJ *et al. Biometrics*. 2005 Mar;61(1):92-105.
7. Li H *et al. JASA*, Aug, 2017
8. Krapohl E *et al. Mol Psychiatry*. 2018 May;23(5):1368-1374.
9. Maier RM *et al. Nat Commun*. 2018 Mar 7;9(1):989.