The Molecular Fountain of Youth:
Can We Identify a Genetic Signature for Human Healthspan?

Morgan E Levine[1], Jennifer A. Ailshire[2], Eileen M. Crimmins[2]

[1]*Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA*

[2]*Davis School of Gerontology, University of Southern California, Los Angeles, CA*

ABSTRACT

A number of studies have attempted to identify alleles influencing human lifespan. However, a better goal may be to identify the underlying genetic mechanisms regulating "healthspan". Using data from the Health and Retirement Study, we attempted to identify SNPs associated with surviving to at least age 90 with no history of disease, versus developing an age-related disease prior to age 70. Genome-wide association results identified 18 SNPs with suggestive association, the most significant of which was an intron variant (rs11928544) located on 3q13.2 in *TMPRSS7*—a gene that may regulate leukocyte telomere length. Finally, based on our GWAS, three polygenic risk scores were created. Using an independent validation sample from HRS, we found that two of these scores were significantly associated with disease risk (P=0.008-0.017). In moving forward, genetic scores for healthspan may help identify at-risk individuals and facilitate our understanding of the molecular mechanisms regulating the aging process.

INTRODUCTION

As far back as the 5th century B.C., mankind has been searching for a way to slow aging. While for explorers like Juan Ponce de León, this may have involved expeditions in search of restorative bodies of water, in the modern era, the path to longevity is being explored using the information in our DNA. Studies using twin data have estimated that normal human lifespan is approximately 25-30% heritable (Finch and Kirkwood 2000; Herskind et al. 1996). However, there is evidence to suggest that the influence of genes on longevity may be even greater when considering survival to extreme ages—such as over one-hundred years (Sebastiani and Perls 2012). This knowledge, coupled with the recent boom in the accessibility of genomic data, has invigorated the search for genes with the potential to increase human lifespan. Over the past decade, there have been a number of studies published, which seek to identify alleles that differ between centenarians, or sometimes super-centenarians, and the general population—for whom average life expectancy is just under 79 years old (Beekman et al. 2013; Deelen et al. 2014; Franceschi et al. 2005; Sebastiani et al. 2012; Suh et al. 2008). However, if our goal is to identify genetic factors which regulate the rate of aging, one has to wonder, are we asking the right question?

In the demographic literature, lifespan has become synonymous with the rate of aging. In human populations, as with many model organisms, after development the mortality rate increases exponentially with age. This observation, termed the Gompertz-Makeham law of mortality, represents an age-dependent component of the death rate, for which the exponential coefficient can be taken to represent the population's "senescent component", or rate of aging (Finch 2007). However, secondary or tertiary medical interventions have the potential to save individuals from death without enacting any influence over the aging process. Under such conditions, population differences or changes in the death rate as a function of age may not represent differences in the rate of biological aging. Therefore, a better goal, when considering the best interests of both individuals and populations, may be to identify the underlying genetic mechanisms regulating "healthspan".

In the U.S., the remaining life expectancy in 2006, among individuals age 65, was predicted to be 17.0 years for males and 19.7 years for females. Unfortunately, I was also predicted that only 47-57% of those years would be spent free from disease (8.1 years for males and 11.3 years for females) and between 26% and 37% would be spent “unable to function” (4.5 years for males and 7.3 years for females) (Crimmins and Beltran-Sanchez 2011). Such discrepancies between lifespan and healthspan are likely to elicit enormous personal, social, and economic burdens. Conversely, if healthspan extension were to keep pace with, or even exceed the rate of lifespan extension (if there is a limit), disease or disability could be postponed to the final year or even months of life, or eliminated entirely (Fries 1980). Unfortunately, this lofty goal, known as “compression of morbidity”, may not be a reality for the majority of human populations. While our life expectancy has skyrocketed over the past sixty years, as a population, we appear to be no closer to living out our lives disability and disease free (Crimmins and Beltran-Sanchez 2011).

Nevertheless, extreme survival without morbidity is not unheard of. Using data on 424 centenarians (aged 97-119 years), Evert et al. (Evert et al. 2003)grouped long-lived individuals into three categories—Survivors (those who survived with a disease), Delayers (those who postponed the onset of disease), and Escapers (those who did not get a disease). While the majority of centenarians either survived after being diagnosed with a disease prior to age 80, or delayed disease onset until at least age 80, surprisingly 32% of the male centenarians and 15% of the female centenarians survived to at least age 100 with no prior diagnosis of any common age-related illness. While most genetic studies of longevity include all three morbidity phenotypes in their case category, Escapers are truly the ones who personify the goal of aging research. Therefore, the goal of this study is to 1) utilize genome-wide association (GWA) analysis to identify candidate SNPs, and 2) create a polygenic risk score (PRS) that represents a genetic signature for healthspan.

METHODS

Discovery and Validation Samples

Participants for our discover sample were part of the 2006 and 2008 waves of the Health and Retirement Study (HRS), a nationally-representative longitudinal study of health and aging in the U.S. Using a similar classification to what was presented in the paper by Evert et al., we grouped participants into two healthspan categories. Cases were defined as those individuals who survived to at least age 90 prior to being diagnosed with any one of six major age-related diseases—cardiovascular disease, diabetes, cancer (excluding skin), stroke, or hypertension. Controls were defined as individuals who were younger than age 80 (at every wave of the HRS) and had been diagnosed with at least one of the six major diseases prior to reaching age 70. These case-control groups were further restricted to include only Non-Hispanic white participants, in order to increase genetic homogeneity and reduce the threat of population stratification in our analysis. Our validation sample also used participants from HRS. These individuals included all genotyped non-Hispanic white participants who were not a part of our discovery sample (cases or controls).

Genotyping and Quality Control

Genotyping in HRS was performed for participants who provided saliva samples and signed consent forms in 2006 and 2008 and was carried-out by the NIH Center for Inherited Disease Research (CIDR) using the Illumina Human Omni-2.5 Quad beadchip, with coverage of approximately 2.5 million single nucleotide polymorphisms (SNPs). Quality control filters were performed by CIDR and the Genetics Coordinating Center of the University of Washington. These filters consisted of removal of: Duplicate SNPs; Missing call rates ≥2%; >4 discordant calls in 423 study duplicates; >1 Mendelian error; Hardy-Weinberg Equilibrium P-values <10-4 in European or African samples; Sex differences in all allelic frequency ≥0.2; and Sex differences in heterozygosity >0.3. As a result, 2,201,371 SNPs remained. However, given our small sample of cases which could inflate P-values for SNPs with small minor allele frequencies (MAFs), we set our MAF cutoff at 0.05, which left us with a total of 1,224,285 SNPs for our analysis. For ELSA, genotyping was performed using the Illumina Omni2.5-8 beadchip and employing the same QC criteria used for HRS.

Principal components analysis (PCA) was conducted for both the HRS and ELSA samples to account for population structure in accordance with the methods described by Patterson et al. This analysis produced sample eigenvectors (EV). A scree plot generated by HRS showed that the 20 components produced by the PCA only accounted for a small fraction of the overall genetic variance (<4%) for the full HRS genetic sample and that most of this was contained within the first two components. Given that our sample was already restricted to Non-Hispanic whites, only the first four eigenvectors were included as covariates in our genetic analyses.

Statistical Analysis

A case-control GWAS was performed using our discovery sample and adjusting for sex and population stratification, to identify candidate single nucleotide polymorphisms (SNPs) that are potentially associated with healthspan. Based on these results, SNPs with P<.05 were clustered into groups by employing the clump command in PLINK. Using this procedure, we grouped SNPs based on linkage disequilibrium ($R^2 > 0.5$) and physical distance (<250 kb), so that only the most significant SNP in each group would be included in our candidate SNP list. PRSs were then calculated for each individual, in order to examine the cumulative effects across multiple SNPs (Wray, Goddard and Visscher 2007). A PRS can be thought of as a measure of 'genetic burden' associated with a phenotype (Wray, Goddard and Visscher 2008). PRSs are generated by summing the minor allele counts (0, 1, 2), weighted by the SNP-specific coefficients from the GWAS. PRS are based on the idea that many variants with small individual effects will not meet genome-wide significance thresholds, yet collectively may have a strong effect (Dudbridge 2013; Wray et al. 2007, 2008). Three PRS were calculated for each individual in our validation sample. The first score included all SNPs that had a significant level of P<0.05 in the GWAS, the second included those with P<0.005 and the third included those with P<0.0005. Finally, Cox proportional hazard models were used to determine whether PRS was associated with healthspan, signified by the timing of first disease incidence.

RESULTS

Sample Characteristics

The mean age at last observation for cases in our discovery sample (n=85) was 93 years (s.d.=2.7; range=90-102). Of these participants, 40 developed a disease, with mean age of first incidence occurring around 92 years. Overall, 12 individuals developed heart disease, 8 developed cancer, 3 developed diabetes, 22 developed hypertension, 8 experienced a stroke, and 4 developed lung disease. Mean age of controls from our discovery sample (n= 5,094) was 67.5 years (s.d.=7.3). All of these individuals developed a disease, with the age of first incidence ranging from 24-70, and a mean of 56.8 years (s.d.=6.2). Approximately 34.5% developed heart disease, 21.1% developed cancer, 33.7% developed diabetes, 80.6% developed hypertensions, 10.6% had a stroke, and 16.6% developed lung disease. Finally, for our validation sample (n=3,787), mean age at last observation was 75.4 (s.d.=11.9). Approximately 55% of the sample developed a disease while enrolled in the study. On average, first disease incidence occurred at age 73.5 (s.d.=5.2). In the overall validation sample, 25.5% of participants developed heart disease, 15.4% developed cancer, 13.5% developed diabetes, 42.5% developed hypertensions, 10.8% had a stroke, and 8.0% developed lung disease

Genome-Wide Association Study

GWAS results for all autosomal SNPs are displayed in Figure 1. Overall, while no set met criteria for genome-wide significance, we found 18 SNPs that met criteria for suggestive significance ($P<10^{-05}$). Based on our QQ-Plot (Figure 1b), we find that overall, there are more SNPs that meet the criteria for suggestive significance than would be expected by chance. Our most significant SNP (rs11928544) had a P-value=$9.38e^{-08}$. This SNP is an intron variant, located on 3q13.2 in the Homo sapiens transmembrane protease, serine 7 (*TMPRSS7*) gene. While to the best of our knowledge, rs11928544 has not been implicated in any disease traits, a nearby SNP (rs16859140), which is also located on 3q13.2 in the *TMPRSS7* gene was found to have a suggestive association with leukocyte telomere length (P=4.90e-06).

Polygenic Risk Score

After clumping SNPs based on linkage disequilibrium, we were left with 32,213 SNPs that were included in our first PRS (PRS1 significant criteria: P<0.05), 3,183 SNPs that were included in our second PRS (PRS2 significant criteria: P<0.005), and 363 SNPs that were included in our third PRS (PRS3 significant criteria: P<0.0005). For our validation sample, participants with more than 10% missing SNP data for each score were excluded. All three PRS were standardized to have a mean of zero and standard deviation of one. Overall, the three PRS ranged from about -4 to 4. The three scores were also moderately correlated. The highest correlation was between PRS1 and PRS2 (r=0.61), followed by the correlation between PRS2 and PRS3 (r=0.55), and finally the correlation between PRS1 and PRS3 (r=0.29).

Results from Cox proportional hazard models are shown in Table 1. After adjusting for age, population stratification, sex, and education, we found that both PRS1 and PRS2 were significantly associated with disease risk in our validation sample. For instance, a one-standard deviation increase in either PRS1 or PRS 2was associated with a 6% decrease in disease risk (PRS1: HR=0.94, P=0.017; PRS2: HR=0.94, P=0.008). Finally, we plotted adjusted Kaplan-Meier curves, as a function of PRS2—individuals two standard deviations below the mean, individuals at the mean, and individuals two standard deviations above the mean. We find that a two standard deviation increase in PRS2 is associated with about a one year postponement in disease incidence.

DISCUSSION

Our study identified a set of SNPs that together were associated with healthspan in an independent sample. The current study moves beyond the one locus approach to examine the cumulative effect of multiple polygenic loci. While GWASs are designed to examine one SNP at a time without considering an individual's other genetic characteristics, it is unlikely that a single mutation accounts for the majority of the genetic risk for most complex traits. Instead, it is more likely that groups of variants, with small individual effects, combine to influence an individual's rate of aging or susceptibility to chronic diseases (Wray et al., 2007; Wray et al., 2008).

The score we identified, based on information from 3,183 SNPs with significance levels of P<0.005 was found to be significantly related to disease incidence. For instance, a one-standard deviation increase in the polygenic score equated to a 6% decrease in risk of one of six major age-related diseases—cardiovascular disease, diabetes, cancer (excluding skin), stroke, or hypertension—as well as a 0.5 year postponement in the timing of first disease.

In moving forward it will be important to examine whether genes pertaining to the SNPs in the score are associated with various biological pathways or processes. It will also be important to examine gene score by environmental effects, to determine if factors (e.g. smoking, obesity, SES) influence the association between the polygenic score and disease risk. For instance, it is possible that the effect of the polygenic score on disease susceptibility is more pronounced for individuals with poor social/behavioral conditions.

Overall, our ability to use genetics to model disease susceptibility has the potential to aid in both primary and secondary prevention strategies, or interventions. For instance, polygenic scores could be used to identify at-risk individuals for inclusion in lifestyle interventions or more intensive screening. Genetic risk assessments are currently based on only a few markers with extremely small effects. However, using polygenic models to understand how variants across the genome collectively influence aging and disease could produce more meaningful estimates. Moreover, examining the associations between these polygenic networks and different environmental conditions will deepen our understanding of the pathways through which genes and environments interact to influence aging and longevity, while also facilitating personalized medical and lifestyle recommendations.

REFERENCES

1. Beekman, M., H. Blanche, M. Perola, et al. 2013. "Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study." *Aging Cell* 12(2):184-193.

2. Crimmins, E.M. and H. Beltran-Sanchez. 2011. "Mortality and Morbidity Trends: Is There Compression of Morbidity?" *Journals of Gerontology Series B-Psychological Sciences and Social Sciences* 66(1):75-86.
3. Deelen, J., M. Beekman, H.W. Uh, et al. 2014. "Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age." *Human Molecular Genetics* 23(16):4420-4432.
4. Dudbridge, F. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." *PLoS Genet* 9(3).
5. Evert, J., E. Lawler, H. Bogan, and T. Perls. 2003. "Morbidity profiles of centenarians: survivors, delayers, and escapers." *J Gerontol A Biol Sci Med Sci* 58(3):232-237.
6. Finch, C. 2007. *The biology of human longevity : inflammation, nutrition, and aging in the evolution of life spans*. Burlington, MA: Academic Press.
7. Finch, C.and T.B.L. Kirkwood. 2000. *Chance, development, and aging*. New York: Oxford University Press.
8. Franceschi, C., F. Olivieri, F. Marchegiani, M. Cardelli, et al. 2005. "Genes involved in immune response/inflammation, IGF1/insulin pathway and response to oxidative stress play a major role in the genetics of human longevity: the lesson of centenarians." *Mech Ageing Dev* 126(2):351-361.
9. Fries, J.F. 1980. "Aging, Natural Death, and the Compression of Morbidity." *New England Journal of Medicine* 303(3):130-135.
10. Herskind, A.M., M. McGue, N.V. Holm, et al. 1996. "The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900." *Hum Genet* 97(3):319-323.
11. Sebastiani, P.and T.T. Perls. 2012. "The genetics of extreme longevity: lessons from the new England centenarian study." *Front Genet* 3:277.
12. Sebastiani, P., N. Solovieff, A.T. Dewan, et al. 2012. "Genetic signatures of exceptional longevity in humans." *PLoS One* 7(1):e29848.
13. Suh, Y., G. Atzmon, M.O. Cho, et al. 2008. "Functionally significant insulin-like growth factor I receptor mutations in centenarians." *Proc Natl Acad Sci U S A* 105(9):3438-3442.
14. Wray, N.R., M.E. Goddard, and P.M. Visscher. 2007. "Prediction of individual genetic risk to disease from genome-wide association studies." *Genome Research* 17(10):1520-1528.
15. Wray, N.R. 2008. "Prediction of individual genetic risk of complex disease." *Current Opinion in Genetics & Development* 18(3):257-263.

Table 1: Results from Cox Proportional Hazard Models (First Incidence of Disease)

	N	Total Person-Years	Hazard Ratio	P-value
PRS1 (Based on 32,213 SNPs w/ P<0.05)	2,847	206,836	0.94	0.017
PRS2 (Based on 3,183 SNPs w/ P<0.005)	3,458	251,092	0.94	0.008
PRS3 (Based on 363 SNPs w/ P<0.0005)	3,787	274,831	0.99	0.494

All models are adjusted for age, population stratification, sex, and education

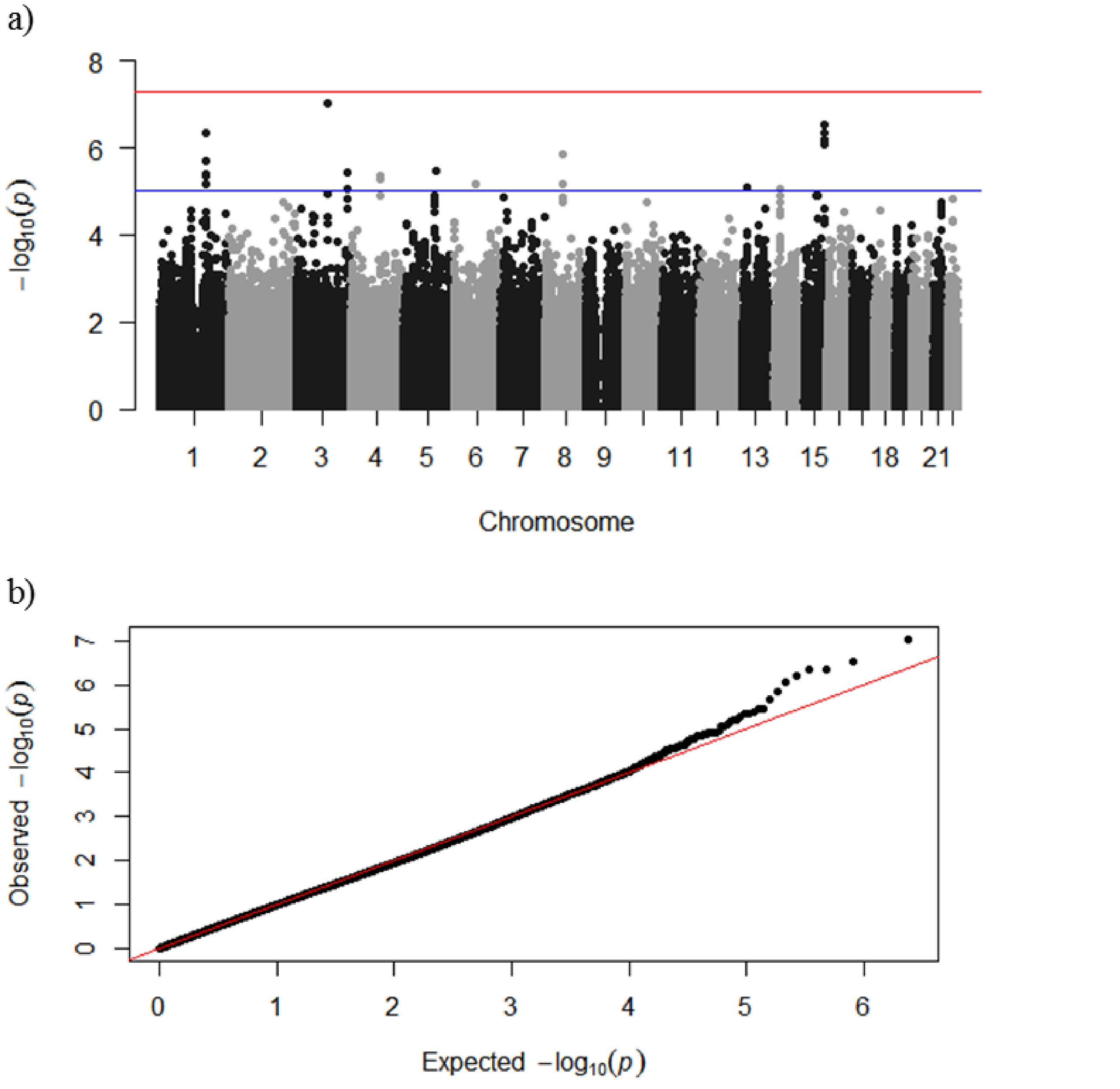

Figure 1: GWAS P-values Shown in a Manhattan Plot (a) and QQ-Plot (b)

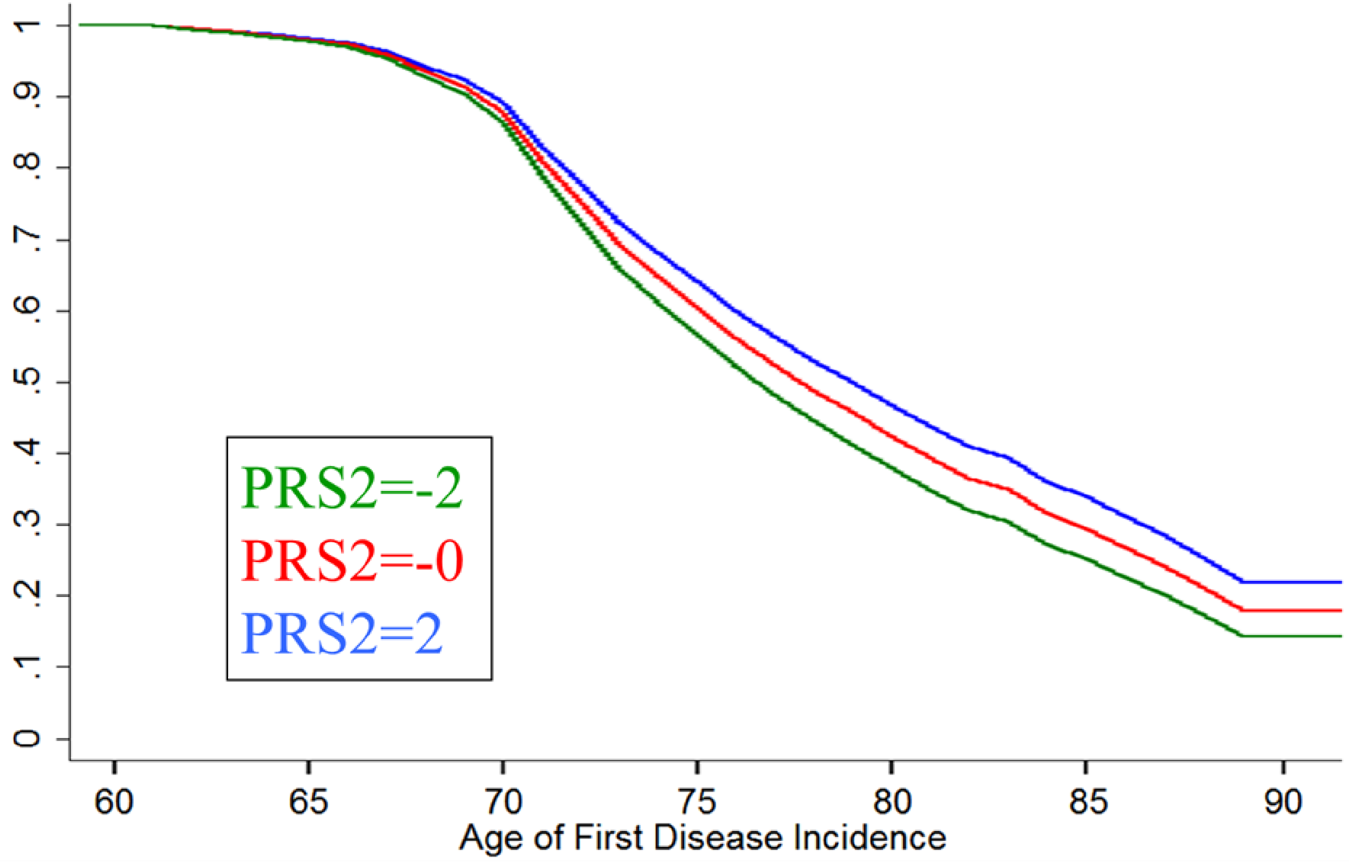

Figure 2: Adjusted Kaplan-Meier Plots by PRS2