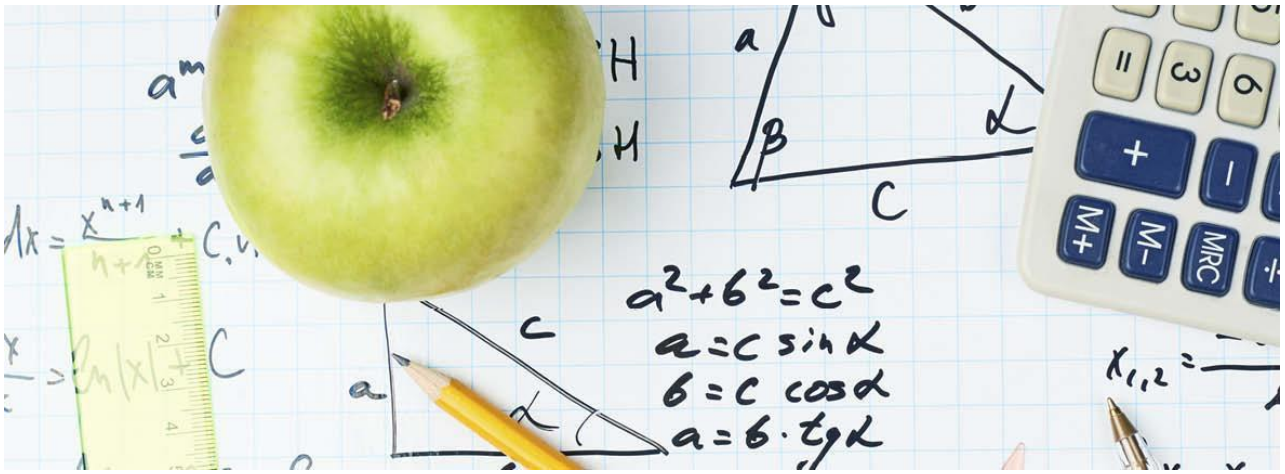
COHERE - Centre of Health Economics Research



Measuring the Effect of the Polygenic Risk Score on the Aging Rate

By

Georgios Effraimidis, COHERE, Department of Business and Economics, University of Southern Denmark
Morgan Levine, UCLA, University of Southern California
Eileen Crimmins, Davis School of Gerontology, University of Southern California

COHERE discussion paper No. 7/2016

FURTHER INFORMATION
Department of Business and Economics
Faculty of Business and Social Sciences
University of Southern Denmark
Campusvej 55,
DK-5230 Odense M Denmark
www.cohere.dk
ISSN 2246-3097

Measuring the Effect of the Polygenic Risk Score on the Aging Rate *

Georgios Effraimidis† Morgan Levine‡ Eileen Crimmins §

September 18, 2016

Abstract

Population aging has emerged as a major demographic trend around the globe. Aging is a process that is determined by millions of genetic factors. The identification of the set of genetic factors that has a ***significant*** role in the aging process is a highly challenging task. This paper studies the association between genetic factors and the aging rate. We first calculate the so-called polygenic risk score (PRS) by following a well-designed algorithm for the selection of the significant single nucleotide polymorphisms (SNPs) and subsequently considering a weighted sum of those significant SNPs. Next, we construct a new mortality model, which allows the aging rate to depend on the PRS. Our statistical analysis is based on a rich dataset from the Health and Retirement Study.

Keywords: Aging rate; Genome-wide association study; Mortality rate; Polygenic risk score.

*We thank seminar participants at the Center of Economic and Social Research for useful comments and suggestions. Effraimidis gratefully acknowledges financial support from the Danish Council for Independent Research via the FP7 Marie Curie Actions COFUND (DFF: 1329-00074A). The funder is not involved in data analysis and preparation of the manuscript.

†Department of Business and Economics, University of Southern Denmark. Email: georgiosnl@gmail.com.
‡Department of Human Genetics, UCLA. Email: melevine@mednet.ucla.edu.
§USC Davis School of Gerontology. Email: crimmin@usc.edu.

1 Introduction

The world's population is growing as well as aging. Population aging has emerged as a major demographic trend around the globe. According to predictions made by demographers, the fraction of elderly dependents will reach rather high levels. Today, three countries (Germany, Italy, and Japan) are characterized as super-aged (i.e., more than 21% of the national population is aged 65 or over). It is expected that Bulgaria, Finland, Greece, and Portugal will become super-aged in the next five years. In the following decade, other European countries, such as Austria, France, Sweden, United Kingdom, as well as non-European countries such as Canada and South Korea, are expected to become super-aged, too.

Population aging is a phenomenon with profound societal implications. Numerous studies have investigated the impact of human aging on global health care expenditure, economic growth of countries, and individual retirement decisions. Thus, detailed analysis of the underlying aging process is becoming an increasingly crucial task, as it can help us acquire a better understanding of the process, and more importantly, enable us to provide more accurate predictions of future mortality.

Aging is a process that is determined by millions of genetic factors. The identification of the set of genetic factors that has a *significant* role in the aging process is a highly challenging task. The main difficulty lies in the fact that individual genetic factors may have no effect, but a combination of these factors could be a strong predictor of the mortality outcome. Given the ultra-high dimension of the problem due to the presence of millions of genetic factors, the finding of appropriate genetic combinations with an effect on the aging process is a very complicated problem.

The main objective of this paper is to build a new model that can help researchers realize what the role of genetic factors in the aging process is. To achieve this objective, we contribute to the existing literature by constructing a mortality model that associates several genetic influences with the mortality outcome. First, we combine a large number of single nucleotide polymorphisms (SNPs) into a single numerical score. The choice of those significant (for mortality) SNPs is made by following a well-designed selection process. In particular, by means of a genome-wide

association study (GWAS) we estimate the effect of each SNP on the mortality outcome. Next, we rank the SNPs using the $p-$value of the corresponding estimated coefficient as a ranking criterion. Finally, we estimate the so-called polygenic risk score (PRS) by considering the weighted sum of the significant SNPs. The current study is the first to make use of a huge number ($\sim$ 30000) of SNPs for the construction of the PRS when mortality is the outcome under consideration. Our new statistical methodology is applied to mortality and genetic data from the Health and Retirement Study.

Second, we introduce a new mortality model that allows the aging rate to be a function of the PRS and analyze the effect of the latter on the former. To the best of our knowledge, this paper is the first one in the field of mortality studies that models the aging rate as individual-specific. The implication of the explicit relationship between aging rate and the PRS is that each individual has his/ her own aging rate, which depends on the PRS- that is, on individual genetic influences. We apply our new model to compare the aging process in two different cohorts: 1911-1920 and 1921-1930.

The remainder of the paper is organized as follows. Section 2 provides a brief discussion of the PRS and its use in different studies in the past. Section 3 focuses on the methodology for estimation of the PRS from our data. Section 4 focuses on nonparametric statistical inference and gives some preliminary results on the relationship between mortality outcomes and the PRS. In section 5, we provide a short review of existing mortality models and discuss the implied aging rate for each of these models. Section 6 introduces a new model for the estimation of the aging rate as a function of the PRS. Section 7 presents the empirical results obtained by using the new model. Section 8 concludes and discusses possible extensions of the current work and, more importantly, different methods for the estimation of the PRS. In the appendix, we have included some technical details of the paper.

2 Polyegenic Risk Score

In recent years, there has been a growing interest in studying how genetic factors can affect the likelihood of a certain phenotype outcome such as disease, death, or health-risky behaviour (e.g.

smoking and/or use of alcohol). It is well-known that it is possible for many of the factors to have no marginal effect on their own, but when they are combined they can provide a good prediction mechanism for the phenotype outcome under consideration. To capture such a (potential) dependence between a set of genetic factors and a phenotype outcome, researchers have developed the concept of PRS. In a nutshell, the PRS is obtained by combining several genetic factors into a single numeric score. To give a more formal definition, the PRS is equal to the sum of trait-associated alleles across many genetic loci, weighted by effect sizes as calculated from a GWAS. Dudbridge (2013) provides a nice introduction to the increasingly important concept of PRS. PRS analysis has been widely performed in numerous studies in order to identify the effect of genetic factors on health-risky behavior. Among many others, Vink et al. (2014) try to investigate whether there are overlapping genetic factors that can explain the well-established association between smoking and the use of alcohol and cannabis. Another example of such studies is Belsky et al. (2013), who build a PRS in order to identify a possible association between genetic factors and progression to heavy smoking, nicotine dependence, and diffculties with cessation of those two health-risky outcomes.

Furthermore, PRS analysis has been extensively applied to understand disease risk. In particular, Aly et al. (2011) calculate a PRS related to prostate cancer and show that it can be a good tool for prediction of the disease. In fact, the authors claim that a proper prediction mechanism for prostate cancer can reduce the number of biopsies in the future. In a well-known study, Cross-Disorder Group of the Psychiatric Genomics Consortium and others (2013) identify genetic contributions to five different diseases: autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia. Stahl et al. (2012) carry out a PRS analysis and develop a new approach to infer which genetic factors play an important role in the onset of rheumatoid arthritis. Additionally, Derks et al. (2012) apply a PRS analysis to examine whether there is a genetic association between quantitative measures of psychosis and schizophrenia.

The study of the association between PRS and mortality is a new topic. Yashin et al. (2012) discuss how the PRS influences individual lifetime by applying several statistical procedures. Specifically, they first choose a set of significant SNPs by using six different methods: (*i*) normal linear

regression, (*ii*) Cox regression, (*iii*) logistic regression; (*iv*) the generalized estimation equation, (*v*) the mixed model and finally (*vi*) the gene frequency method. After identifying the set of significant SNPs for each applied method from an initial 550K set, they also obtain 27 SNPs from the intersection of those different six estimation methods. In the second step, for each of these 6 + 1 methods (the seventh method uses the SNPs from the intersection of the six different methods), they calculate the PRS and estimate the effect of PRS on life span by employing linear regression. Trying also to determine the channel through which those SNPs affect the mortality outcome, they find out that almost half of the genes related to the 27 SNPs play a role in cancer development and one third of the 27 SNPs play a significant role in brain activities that are highly relevant to brain aging.

In an attempt to (partially) explain exceptional longevity, Sebastiani et al. (2012) focus on 801 centenarians. They conduct a PRS analysis and confirm that the genetic contribution is largest for the oldest ages. By means of Bayesian as well as frequentist statistical approaches, they first rank SNPs based on their degree of association with mortality outcomes. The association analysis makes use of almost 250000 SNPs. Next, they use a Bayesian classification model and divide their sample into a training sample and a test sample in order to choose the most significant (for the lifespan) SNPs. The strength of this approach is that it does not impose any restriction on the number of SNPs that play a signifiant role for the lifespan. The proposed method yields 281 predictive SNPs- that is, those SNPs give a high posterior probability of exceptional longevity.

Walter et al. (2012) begin with a set of 43 SNPs. They finally select five SNPs as predictors for mortality. However, they claim that the predictive gains due to the inclusion of those five SNPs is rather low as long as other socioeconomic characteristics are taken into account for mortality prediction.

Ganna et al. (2013) have an initial set of 5488 SNPs from 1128 different studies. By following a multi-step procedure, they end up with a set of 707 SNPs and estimate the resulting PRS as the sum of the corresponding risk alleles. Next, they use a Cox proportional hazard model to make statistical inference regarding the association between (*i*) PRS and age at death and (*ii*) PRS and age at incidence of one the following nine major diseases: coronary heart disease, stroke, heart

failure, diabetes, dementia, as well as lung, breast, colon, and prostate cancers. Furthermore, in order to prove that their analysis is robust to any possible mispecificaton concerning the calculation of the PRS, they also develop 17 other PRSs by considering different number of SNPs for each different PRS.

We have also carried out a little study to investigate whether there are common significant SNPs among the four aforementioned papers, namely: Yashin et al. (2012), Sebastiani et al. (2012), Walter et al. (2012), and Ganna et al. (2013). Our finding is that there is only one single common SNP between the second and the fourth papers. *rs*2075650 is the SNP that is present in both studies.

Finally, Hamad and Rehkopf (2015) are the first to try to establish a causal relationship between telomere length and health status. Provided that these two variables are affected by unobserved factors, they use genetic factors as an instrument in order to identify the effect of telomere length on various health outcomes. Specifically, they estimate the PRS based on a set of significant SNPs and use that variable as an instrument for the telomere length. They also apply a Cox proportional hazard model to determine the relationship between the PRS and mortality outcomes.

3 Our Polyegenic Risk Score

Our analysis makes use of data from the Health and Retirement Study. That study has genotyped individuals in two different waves: 2006 and 2008. It has also collected phenotype information during the period 1992-2012. In our analysis, we end up with 9480 individuals of European ancestry for whom we have information about their (possibly censored) mortality outcome, demographic characteristics (e.g., gender and education) and genetic profile. The individual genetic profile in this study consists of almost 1.2 million SNPs. Note that for the estimation of the PRS we use the entire sample (i.e., 9480 individuals), whereas for the analysis of the mortality outcomes we focus only on two cohorts due to heavy censoring of the younger cohorts. Another reason for working with two cohorts is to compare the underlying aging processes. As will be shown later, those two cohorts are different in terms of the distribution of the PRS as well as the aging process.

In contrast to previous studies, which determine a small set of significant SNPs and then

estimate the PRS, our estimation method for the PRS makes use of a large number of significant SNPs. We first conduct a GWAS that examines the association between each SNP and a given phenotype, which is mortality in the present study. This allows us to estimate the marginal effect of each SNP and obtain the respective weights. In summary, the GWAS involved running about 1.2 million logistic regression models, each with the mortality indicator (i.e., whether death is observed) as the dependent variable , the SNP as the explanatory variable, and adjusting for covariates (sex, age, etc.).

In mathematical notation, let D_i be an indicator that is equal to 1 if death is observed for individual i and 0 otherwise, where $i \in \{1, 2, \ldots, 9480\}$. Also, let R_i denote the covariate set for individual i without including any of the individual SNPs in that set. SNPs are coded as 0,1,or 2, depending upon the participant's number of minor alleles for that SNP. Formally, to estimate the effect of an individual SNP_i^j on the mortality outcome indicator D_i, we consider the following logistic regression for each $j \in J$

$$\ln \frac{\mathrm{P}(D_i = 1 | R_i,\ SNP_i^j)}{1 - \mathrm{P}(D_i = 1 | R_i, SNP_i^j)} = \beta^j + \gamma^j SNP_i^j + \delta^j R_i, \qquad (1)$$

where $J := \{1,\ 2,\ \ldots,\ J\}$ and the cardinality of J (i.e., J) is the initial number of SNPs under consideration. In particular, $J = 1271442$.

Therefore, for each j we get the estimates $\hat{\beta^j}$, $\hat{\gamma^j}$, and $\hat{\delta^j}$. The main problem in this step is to determine the set of the most significant SNPs associated with death. The selection process begins by defining a p-value cut-off point. We choose the, rather conventional in the literature, value of 0.05. Based on the $p-$value of the estimated $\hat{\gamma^j}$, we rank the SNPs in ascending order. The number of SNPs with $p-$value of the estimated coefficient $\hat{\gamma^j}$smaller than 0.05 is close to 65000. After "pruning", meaning clumping SNPs based on their correlations and distance, the number of significant SNPs, S, is close to 40000, specifically $S = 38219$. In summary, empirical estimates of Linkage Disequilibrium are used in order to group the SNPs that have $p-$value smaller than 0.05. To carry out the grouping procedure, a distance of 250 kilobases and an R^2 threshold of 0.50 are adopted. See Levine et al. (2014) for the selection procedure of significant SNPs in case the depressive symptoms is the outcome of interest.

After having identified the set of singificant SNPs we proceed with the estimation of the PRS for each individual i. Specifically, we have for each $i \in \{1,2,\ldots,9480\}$

$$PRS_i = \sum_{j=1}^{S} SNP_i^j \hat{\gamma}^j = SNP_{i1}\hat{\gamma}^1 + SNP_{i2}\hat{\gamma}^2 + \ldots + SNP_{iS}\hat{\gamma}^S. \quad (2)$$

With a slight abuse of the notation, the subscript j in the above equation refers only to the significant SNPs. Note that, in contrast to existing studies with focus on the association between the PRS and the mortality outcome, we weight each SNP j with the estimated coefficient $\hat{\gamma}^j$. The reason why we choose to adopt such a weighting scheme is that each SNP makes its own individual contribution to the mortality outcome.

Figure 2 plots the local linear estimator of the density function of the variable PRS.

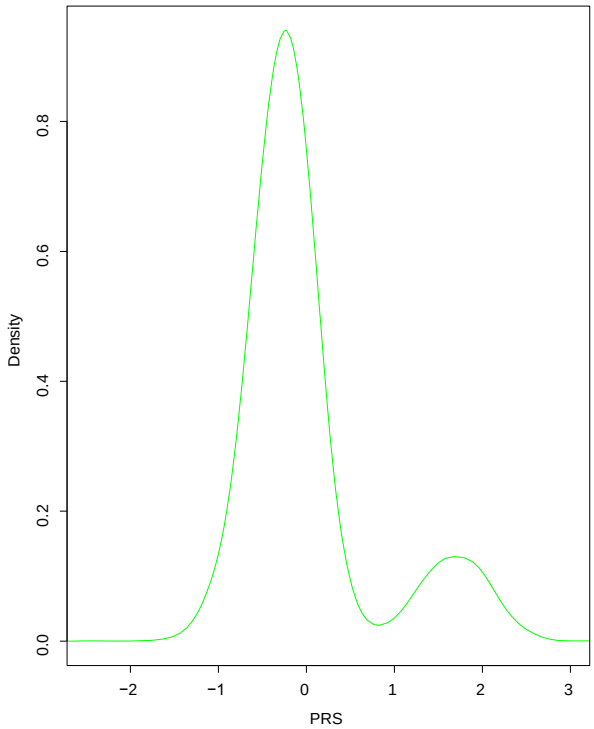


Figure 1: Local linear point estimator of the density function of the PRS. The estimator makes use of the Gaussian kernel, and the bandwidth is chosen according to the Silverman Rule-of-Thumb.

The distribution is slightly right skewed with the skew to be to the right, and the mean (0) is

just right of the median (−0.19). Additionally, the distribution of the PRS is unimodal with the median to the right of the unique mode.

Figure 2 illustrates the relationship between realized lifetime versus PRS. Moreover, it depicts the local linear estimator of the lifetime expectancy as a function of the PRS. Only uncensored observations have been used for the creation of the plot and the graph. Note that our plot uses only positive values of the PRS as negative values of the PRS are observed only for censored observations.

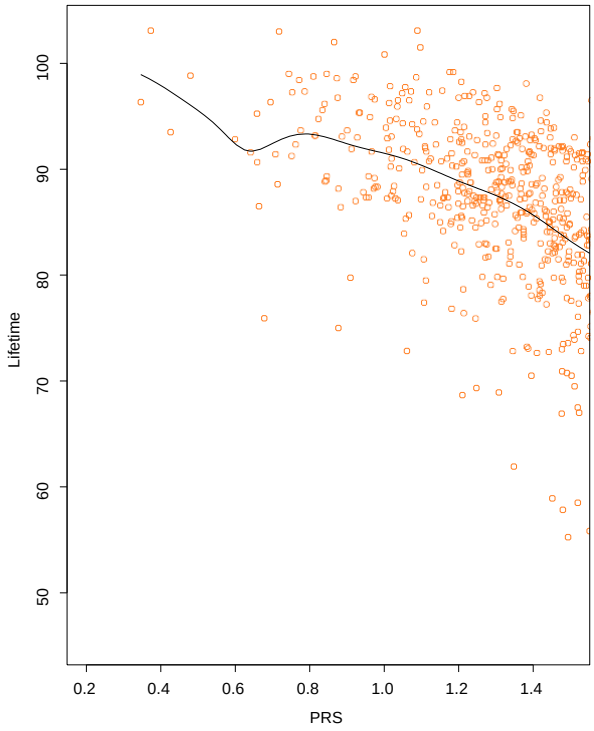


Figure 2: Plot of realized lifetime versus PRS by using only the uncensored observations from the whole dataset. The solid line represents estimated lifetime expectancy; it is obtained by local linear regression, where the Gaussian kernel is employed. The bandwidth is chosen according to the Silverman Rule-of-Thumb.

4 Nonparametric Association between Polygenic Risk Score and Mortality Outcomes

Recall that for the analysis of the relationship between the PRS and mortality outcomes, we consider for comparison purposes only two cohorts: 1911-1920 and 1921-1930. Table 1 gives information about basic statistical measures of the variable PRS for the cohorts 1911-1920 and 1921-1930.

Cohort (Size in parentheses)	Mean	Median	Standard Deviation	Min	Max
1911-1920. (500)	0.28	0.76	1.14	-1.74	2.53
1921-1930 (1944)	0.19	-0.39	0.99	-1.84	2.85

Table 1: Descriptive analysis of the PRS for the cohorts 1911-1920 and 1921-1930.

The PRS take a range of values from -1.74 to 2.53 for the cohort $1911-1920$ and -1.84 to 2.85 for the cohort 1921-1930 The expected value of PRS is 0.28 for the first cohort and 0.19 for the second cohort. Additionally, the median PRS in our two cohorts is 1.14 and 0.99, respectively.

Figure 3 depicts the nonparametric density of the PRS for the cohorts 1911-1920 and 1921-1930.

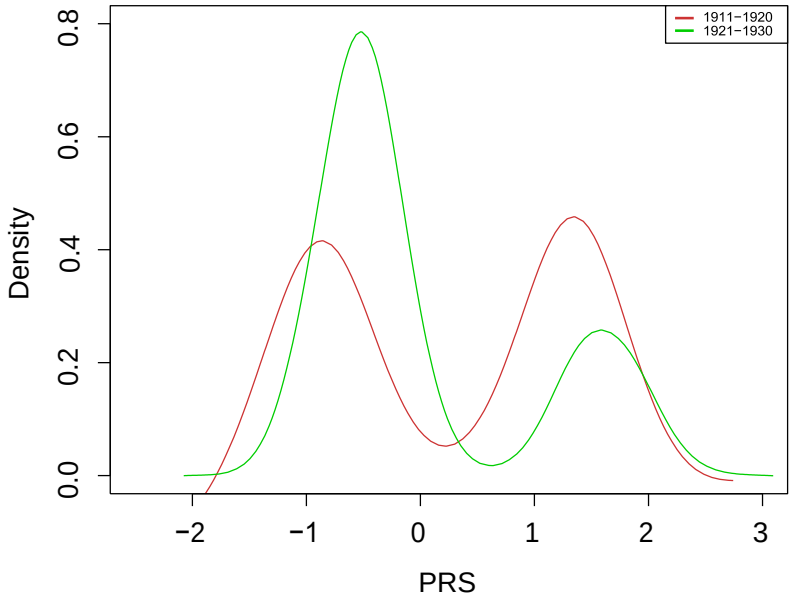


Figure 3: Local linear point estimator of the density function of the PRS for the cohorts 1911-1920 (brown line) and 1921-1930 (green line). The Gaussian kernel is employed and the bandwidth is chosen according to the Silverman Rule-of-Thumb.

Looking at the above figure, we notice two peaks for each probability density function- that is, the underlying densities are bimodal. However, the difference between the two densities is that the density of the first cohort is more symmetric in the sense that the values of the density evaluated for the two modes are almost the same. On the other hand, the PRS distribution of the second cohort has "more" density concentrated to the right of 0, with the value of density evaluated at the first mode being larger than the value of the density evaluated for the second mode.

Table 2 and Table 3 provide descriptive statistics for one of our main variables, that is, mortality outcome. In the cohort 1911-1920, about half of the mortality outcomes are censored, while in the cohort 1921-1930 the censoring is more pronounced given that about 3 in 4 mortality outcomes, 75%, are censored.

Case (Size in parentheses)	Mean	Median	Standard Deviation	Min	Max
Uncensored obs. (262)	92.56	92.29	2.23	86.58	99.17
Censored obs. (238)	96.21	96	2.79	94	103

Table 2: Descriptive analysis of mortality outcomes for individuals born between 1911 and 1920

Case (Size in parentheses)	Mean	Median	Standard Deviation	Min	Max
Uncensored obs. (499)	84.42	84.50	3.13	75.58	91.58
Censored obs. (1445)	87.66	87	2.72	84	93

Table 3: Descriptive analysis of mortality outcomes for individuals born between 1921 and 1930

In order to reach a first understanding of the association between PRS and mortality outcomes, we nonparametrically estimate the survival function for two subpopulations of the cohorts 1911-1920 and 1921-1930. Specifically, the first subpopulation refers to all individuals with a PRS between 1 and 1.5, whereas the second subpopulation refers to all individuals with a PRS between 1.5 and 2. Figure 3 (cohort 1911-1920) and Figure 4 (cohort 1921-1930) depict the Kaplan-Meier estimates of the survival curves for each of the two subpopulations. The red line is for the first subpopulation and the blue line is for the second subpopulation.

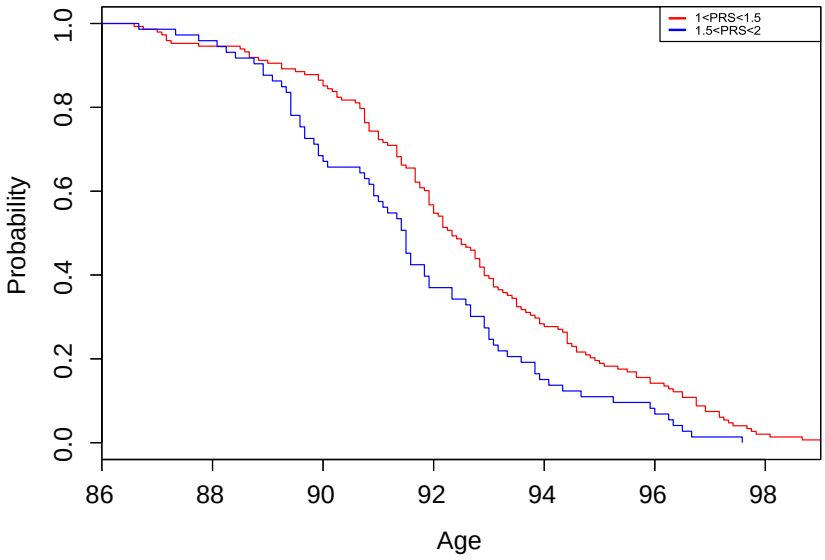


Figure 4: Plot of the Kaplan-Meier survival curves for two subpopulations of the cohort 1911-1920: individuals with PRS between 1 and 1.5 (red line), individuals with a PRS between 1.5 and 2 (blue line).

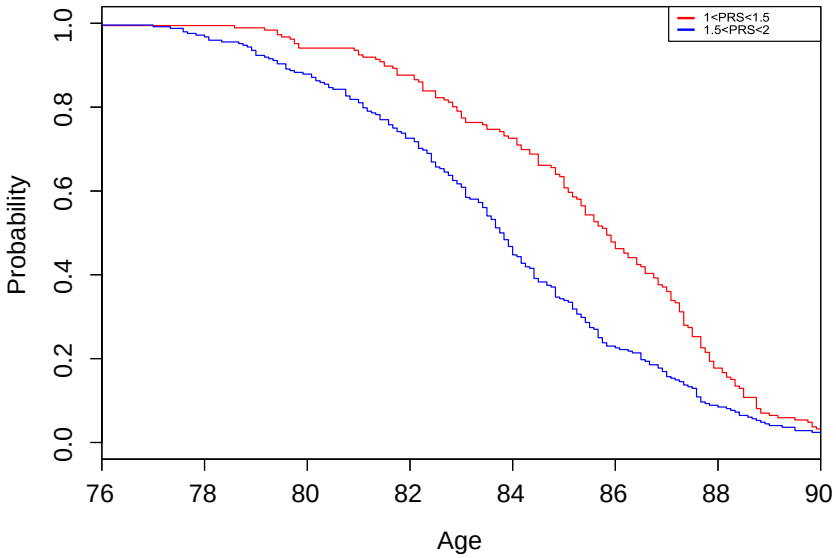


Figure 5: Plot of the Kaplan-Meier survival curves for two subpopulations of the cohort 1921-1930: individuals with PRS between 1 and 1.5 (red line), individuals with a PRS between 1.5 and 2 (blue line).

Looking at the Figures 4 and 5, we can conclude that for any given age the probability of survival for the first subpopulation (small values of PRS) is larger than the probability of survival for the second subpopulation (large values of PRS). In other words, the higher the PRS the higher the likelihood of death at each age. Consequently, as can be expected from these two figures, the life expectancy will be a strictly decreasing function of the PRS. This is confirmed by Figure 6, which contains two plots (PRS versus realized lifetime) and two estimated life expectancies as a function of the PRS. The life expectancies have been estimated by employing the local linear estimator. For the two plots as well as the nonparametric estimators, only uncensored observations have been used.

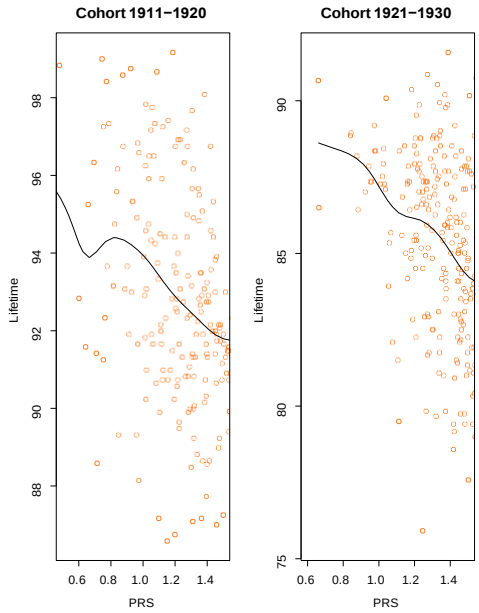


Figure 6: Plot of realized lifetime versus PRS for only the uncensored observations from two different cohorts: 1911-1920 and 1921-1930. The solid lines represent estimated lifetime expectancy; they are obtained by local linear regression where the Gaussian kernel is employed. The bandwidth is chosen according to the Silverman Rule-of-Thumb.

In the next section, we introduce a new mortality model which tries to give an in-depth explanation of the expected lifetime differences due to different values of the PRS.

5 Aging Rate

The mortality rate is the focal point of model building in mortality analysis. In the sequel, we will use the generic symbol m to represent the individual mortality rate. In particular, let X be the stochastic variable that represents the age at death. The realization of that stochastic variable is denoted by x. The mortality rate, $m(x)$, at age x is defined as follows

$$m(x) := \lim_{dt \to 0} (dt)^{-1} \, \mathrm{P}(x \leq X < x + dx | X \geq x). \tag{3}$$

In words, the quantity $m(x)$ gives the instantaneous rate of death at age x given that the individual has survived until the age $x-$.

One of the most fundamental concepts in mortality studies is the aging rate, which determines

the evolution of the mortality rate over the lifespan. By definition, the aging rate equals to the growth rate of the individual mortality rate. In other words, the aging rate equals the derivative of the logarithm of the mortality rate with respect to age. Denote by $ar(x)$ the aging rate at age x. In mathematical notation we have:

$$ar(x) := \frac{\partial \ln m(x)}{\partial x} = \frac{\frac{\partial m(x)}{\partial x}}{m(x)}, \tag{4}$$

where the symbol ∂ refers to the partial derivative.

In the analysis of mortality data, the vast majority of researchers use one of the following three mortality models: (*i*) Cox regression, (*ii*) the Gompertz model, and (*iii*) the model developed by Vaupel et al. (1979). There are a couple of differences between these three models. For instance, the first one is semiparametric, as the baseline hazard is not parameterized. Moreover, the difference between the model introduced by Vaupel et al. (1979) and the other two is that the former includes an extra term, which accounts for individual unobserved variables that presumably affect the mortality rate. As we will explain below, in spite of these major differences the implication of the three models for the aging rate of any given cohort is the same: all individuals have identical aging rates. However, such an implication is not plausible provided that the existence of genetic influences will possibly result in different aging rates among individuals. Therefore, each individual, based on genetic influences, will have his/ her own pace of mortality increase. Below, we first give an overview of the three most popular mortality models employed by empiricists and we discuss why all of them yield an identical aging rate for all individuals of a given cohort.

Let y be a vector of observed socioeconomic, demographic, or genetic characteristics that affect the mortality rate. The well-known Cox regression model has the following structure:

$$m_{Cox}(x|y) = x(x)\exp(ky), \tag{5}$$

where the function $x(x)$ is positive and is left unspecified. Of course, $\ln m_{Cox}(x|y) = \ln x(x) + ky$. Assuming that the function $x(y)$ is differentiable, the aging rate corresponding to this mortality model is equal to $ar_{Cox}(x) = \frac{\partial \ln m_{Cox}(x|y)}{\partial x} = \frac{\partial \ln x(x)}{\partial x}$.

On the other hand, the popular Gompertz model assumes that the mortality rate exponentially increases with age. Particularly, for some positive parameters a and b

$$m_G(x|y) = a\exp(bx)\exp(ky). \tag{6}$$

In view of the definition of the aging rate, we have $\ln m_G(x|y) = \ln a + bx + ky$. It then follows that $ar_G(x|y) = \frac{\partial \ln a+bx+ky}{\partial x} = b$.

One major drawback of the two models above is the absence of unobserved (to the researcher) factors that possibly affect the mortality rate. In their seminal paper, Vaupel et al. (1979), introduce a new mortality model that accounts for those genetic/environmental influences. Specifically, let Z be a nonnegative random variable that captures those (unobserved) characteristics. The mortality rate can now be expressed as follows

$$m_V(x|y,Z) = a\exp(bx)\exp(ky)Z. \tag{7}$$

The variable Z is commonly referred to as frailty and accordingly the above hazard model is called the frailty model. A widely used assumption for the distribution of Z is that it is gamma distributed. Namely, the density f_Z of Z is equal to

$$f_Z(z) = \frac{1}{\frac{1}{k^k}\Gamma(k)} v^{k-1}\exp(-kv),\ z>0, k>0,$$

where the Γ is computed by $\Gamma(k) = \int_0^\infty \omega^{k-1}\exp(-\omega)d\omega$. The fact that the scale parameter is equal to k is just a normalization. One consequence of modeling the distribution of Z as gamma is that the resulting unconditional (on Z) survival functional has a closed-form expression. Another rationale behind this assumption is that it generates an observed mortality rate (i.e., mortality rate not conditional on Z), which levels off at high ages. In subsection 7.3, we discuss a combination of this model and our model and show that after using genetic data in our analysis, we do not need to account for unobserved heterogeneity in the new model that we study in this article. Regarding the aging rate found by using the mortality model (7), we have $\ln m_V(x|y,Z) = \ln a + bx + ky + \ln Z$.

The latter gives $ar_V(x|y) = \frac{\partial \ln a+bx+ky+\ln Z}{\partial x} = b$.

In conclusion, although there are some fundamental differences between models (5), (6), and (7), the resulting aging rate has the same property for all these models. Specifically, all individuals who belong to a certain cohort are subject to the same aging process. In fact, the aging rate is time-varying for model (5), whereas it has a fixed value (i.e., b) for the mortality models (6) and (7). However, the time dependence of the first aging rate is the same for all individuals of a given cohort, and thus the aging rate is identical across all individuals.

6 Aging Rate as Function of the Polygenic Risk Score

Given the presence of genetic characteristics in the underlying mortality process, we firmly believe that the aging rate will depend on (some of) those characteristics. Our aspiration is to develop a new statistical model that will explicitly link the genetic characteristics to the aging rate. Our major contribution with respect to the study of the association between the PRS and the aging rate is to construct a model that will yield an individual-specific aging rate, which is not true for the mortality models (5), (6) and (7).

To achieve this, we specify the aging rate as a function of the PRS. We write down the mortality rate $m(x|PRS)$ as follows

$$m(x|PRS) = a\exp(b\exp(c * PRS)x), \tag{8}$$

where the parameter c captures the effect of the PRS on the aging rate. In case $c = 0$ (i.e., the polygenic score has no effect on the aging rate), $m(x) = a\exp(bx)$, which is the standard Gompertz model. Provided that we control for genetic factors (through the PRS), we temporarily choose, in contrast to Vaupel et al. (1979), not to add an unobserved frailty term. Later, we will also study a specification with unobserved factors and as we will discuss our estimation findings suggest that such unobserved heterogeneity does not exist once we adjust for genetic characteristics through the PRS.

The aging rate for the new model (8) can be calculated as follows

$$a(x|PRS) = \frac{\partial \ln m(x|PRS)}{\partial x} = \frac{\partial(\ln(a) + b\exp(c * PRS)x)}{\partial x} = b\exp(c * PRS). \quad (9)$$

Looking at the above equation, we understand that the aging rate is individual-specific since it depends on the PRS through the term $\exp(c * PRS)$. Clearly, for $c > 0$ ($c < 0$) the PRS has a positive (negative) effect on the underlying aging rate. In view of graphs 4, 5, and 6, we expect that the value of c will be strictly positive. For $c > 0$, individuals age faster and consequently live shorter on average. This is explained by the fact that the instantaneous probability of death at each age, given survival up to that age increases with the PRS. Consequently, individuals will have a lower life expectancy. In the appendix, we provide closed form expressions for the survival as well as density functions. We also conduct a Monte Carlo experiment to show that the estimation algorithm works.

An alternative model for assessing the effect of the PRS on mortality could be

$$m(x|PRS) = a\exp(b * PRS * x).$$

In this model, the aging rate is equal to $\frac{\partial \ln m(x|PRS)}{\partial x} = \frac{\partial(\ln(a)+(c*PRS)x)}{\partial x} = c * PRS$. Note that we do not explicitly include parameter c, as it is absorbed the parameter b. This model is a bit problematic as it implies that the ratio of the aging rates of two individuals with polygenic scores PRS_1 and PRS_2, respectively, is equal to $\frac{PRS_1}{PRS_2}$- that is, the ratio does not depend on c. Finally, another limitation of the above model is that it does not include the Gomperrtz model as a special case. For any nonzero value of the product bc, the aging rate at the individual level will depend on the PRS. In the next section, we focus exclusively on model (8) and use that model to investigate the aging process for two different cohorts.

7 Empirical Results

In this section, we present the estimates for the parameters of model (8) for each of the two cohorts: 1911-1920 and 1921-1930.

7.1 Cohort 1911-1920

Using the 500 observations from the cohort 1911-1920, we get the following estimates (with standard errors in parenthesis)

$$\hat{a} = 1.147e-17 \ (1.864e-17),$$

$$\hat{b} = 0.377 \ (0.0171),$$

$$\hat{c} = 0.0535 \ (0.0033)$$

Straightforward calculations give that the $t-$ values for the above estimates are: 0.6157, 22.0181, 16.6082. In view of those numbers, we can conclude that we have strong statistical evidence that parameter c is significantly different from zero. Equivalently, the PRS has a significant effect on the aging rate. Figures 7 and 8 plot point estimates of the life expectancy as a function of the PRS and point estimates of the aging rate as a function of the PRS, respectively.

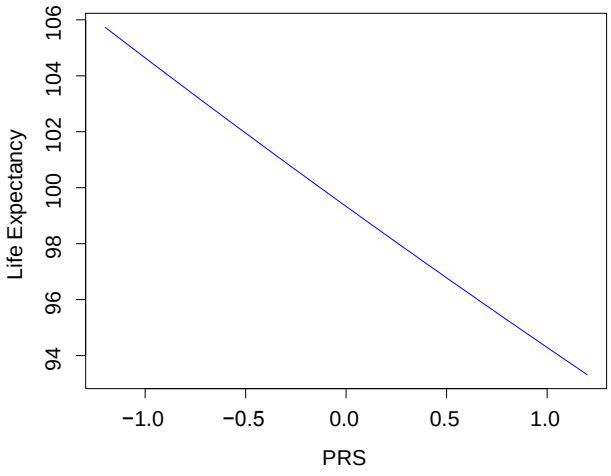


Figure 7: The life expectancy plotted as a function of the PRS for the cohort 1911-1920.

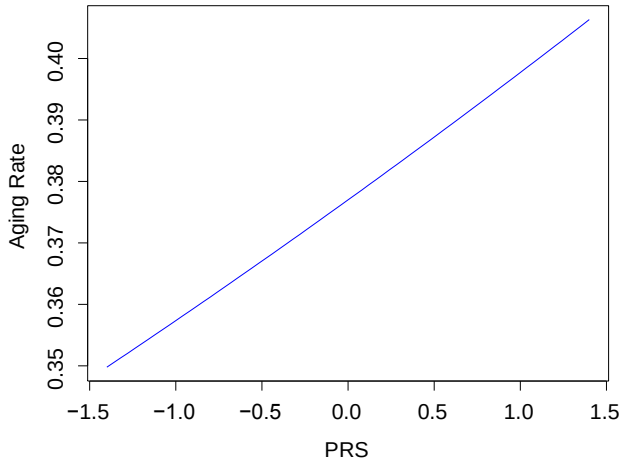


Figure 8: The aging rate plotted as a function of the PRS for the cohort 1911-1920.

7.2 Cohort 1921-1930

Using the 1944 observations, we get the following estimates (with standard errors in parenthesis)

$$\hat{a} = 2.446e-15\ (1.881e-15),$$

$$\hat{b} = 0.337\ (0.0085),$$

$$\hat{c} = 0.071\ (0.0024).$$

Working in an analogous manner as in the cohort 1911-1920, we first calculate the following $t-$ values for the three estimates: $1.1621, 33.9635, 29.5596$. Therefore, for this cohort too, we find out that there is significant relationship between the PRS and the aging rate. 9 and 10 plot point estimates of life expectancy as a function of the PRS and point estimates of the aging rate as a function of the PRS, respectively.

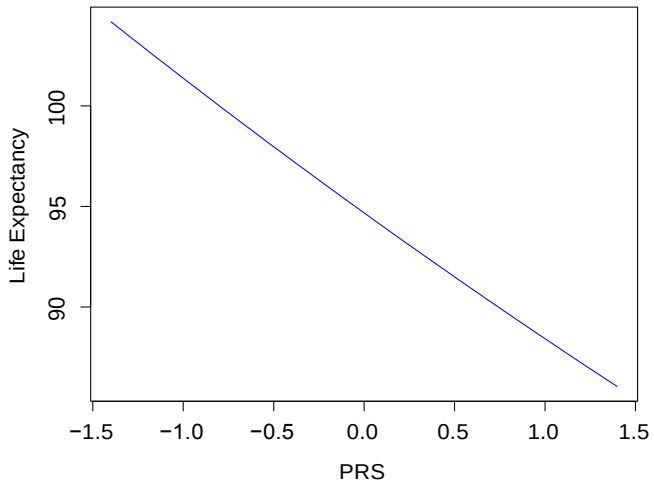


Figure 9: The life expectancy plotted as a function of the PRS for the cohort 1921-1930.

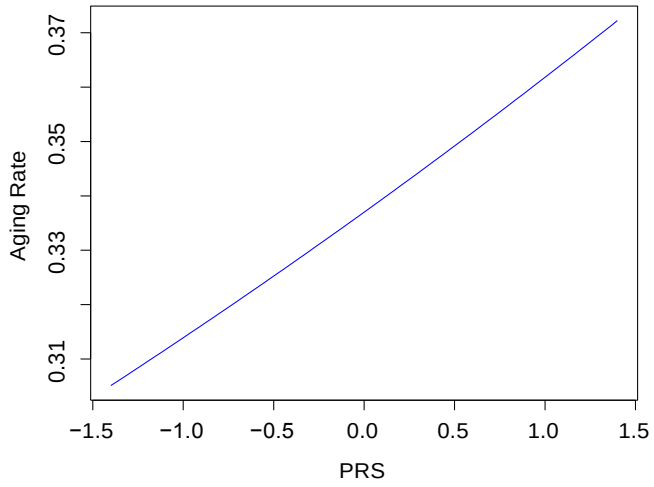


Figure 10: The aging rate plotted as a function of the PRS for the cohort 1921-1930.

Figures 7 and 9 depict lifetime expectancy as a function of the PRS. Our conclusion is that the average lifespan in the cohort 1911-1920 is larger than its counterpart for the cohort 1921-1930. In fact, the average lifespan discrepancy is more pronounced for low values of the PRS. Figures 8 and 10 graphically show the relationship between the aging rate and the PRS. One key conclusion that arises is that for any given value of the PRS, individuals of the cohort 1911-1920 age faster than individuals of the cohort 1921-1930. At first glance, the following statement may

seem contradictory: individuals born between 1911 and 1920 age faster than individuals 1921-1930, yet they live longer. An explanation of such puzzling observation is given by the (estimated) values of a. We notice that the estimated value of this parameter is larger for the cohort 1921-1930. This essentially implies that the initial mortality conditions of individuals born during 1911-1920 are better than these ones of individuals born during the period 1921-1930. In other words, individuals in the former cohort have better health conditions than individuals in the latter cohort. Hence, compared to the cohort 1911-1920, the average lifespan loss due to initial health conditions that individuals of the cohort 1921-1930 experience is smaller than the average lifespan attributed to the aging process. Finally, note that for both cohorts the estimates for parameters a and b in our model are substantially different than their counterparts, which are usually obtained using the standard Gompertz model. This fact is hardly surprising provided that in our model we control for genetic factors, and thus our parameter estimates are not comparable to the estimates that will be obtained with the Gompertz model.

7.3 Mortality Model with PRS and Unobserved Heterogeneity

Recall that the new model (8) developed in this paper explicitly allows the aging rate to depend on genetic factors through the PRS. Of course, someone could claim that there is some unobserved heterogeneity left (mainly environmental factors) that is not captured by this specification. In this subsection, we also briefly discuss the following hazard rate model

$$m(x|PRS, Z) = a\exp\left(b\exp(c * PRS)x\right) Z. \tag{10}$$

The above hazard model combines features from two models: the (7) developed by Vaupel et al. (1979) and our new model (8). On the one hand, the variable Z represents unobserved factors that affect the mortality rate; a statistical feature that is present in (7). On the other hand, the mortality equation (10) allows the aging rate to explicitly depend on the PRS as is the case in the mortality equation (8). As in (7), the random variable Z captures some unobserved characteristics and follows a gamma distribution with a shape parameter equal to κ and a scale parameter equal to $\frac{1}{\kappa}$ (the latter is an innocuous normalization). For that distribution, we have that $E(Z) = 1$

and $Var(Z) = \frac{1}{\kappa}$. Let $q(x|PRS) = a\exp(b\exp(c * PRS)x)$ and $Q(x|PRS) = \int_0^x q(\omega|PRS)d\omega$. By using simple Laplace Transform arguments, the survival function is written as

$$S(x|PRS) = (1 + \underset{\kappa}{1} Q(x|PRS))^{-\kappa}. \tag{11}$$

The density function is expressed as

$$f(x|PRS) = q(x|PRS)(1 + \underset{\kappa}{1} Q(x|PRS))^{-\kappa-1}. \tag{12}$$

Our estimation results (available upon request) show that the parameter κ tends to infinity, which in turn implies that the variance of the stochastic variable Z goes to zero. Hence, the distribution of Z converges to a degenerate distribution concentrated on 1. The interpretation of this limiting result is that after having controlled for genetic factors through the PRS, there is no unobserved heterogenity left among individuals.

8 Conclusions

In this paper, we focus on identifying of the relationship between several genetic factors and the aging rate. By using genetic information and mortality outcomes for two cohorts, 1911-1920 and 1921-1930, we estimate the effect of the polygenic risk score (PRS) on the aging rate and conclude that this effect is significantly different from zero. We also provide point estimates of life expectancy as a function of the PRS.

The first step is to calculate the PRS after having conducted a genome-wide association study. In that step, we estimate the effect of each single nucleotide polymorphism (SNP) by means of a univariate logistic regression after controlling for different demographic characteristics. The number of SNPs under consideration is about 1.2 million. After having estimated the respective coefficients, we implement a certain algorithm to determine the most significant SNPs. The selection criterion being employed in this algorithm is the corresponding $p-$ value obtained for each individual estimated coefficient from the univariate logistic regressions. Particularly, we choose to

include in the significant set all the SNPs whose estimated coeffcient has a $p-$ value larger than 0.05. Next, we calculate the PRS by computing the linear combination of all significant SNPs, where the weights are the estimated coefficients from the logistic regression.

The second step is concerned with the construction of a new mortality equation, which models the aging rate as a function of the PRS. Our model generalizes different existing mortality models by allowing dependence between the aging rate and the PRS. In case the effect of the PRS on the mortality rate is zero, the model reduces to the standard Gompertz model. Finally, having an explicit expression for the mortality rate as a function of the PRS, we obtain a closed-form expression for the probability density function of lifetime. Hence, for each value of the PRS we also obtain an estimate for life expectancy.

We use our new mortality model to study the aging process of two different cohorts: 1911-1920 and 1921-1930. One interesting finding is that despite the fact that the first cohort ages faster than the second cohort, it experiences a longer lifetime expectancy due to better mortality conditions at birth.

In addition, we study a generalized version of our mortality model by allowing the presence of unobserved heterogeneity even after controlling for genetic factors. Our estimation results (available upon request) show that the variance of the stochastic variable that captures those unobserved influences goes to zero. This finding suggests that after adjusting for genetic factors through the PRS, there is no unobserved heterogeneity left .

There are numerous ways in which the analysis of this paper can be extended. It would be interesting to estimate the PRS in the first step by making use of weights obtained by single Cox regressions. That is, instead of using logistic regressions in the first step, we could estimate the effects of the individual SNPs by employing the following univariate Cox regression for each SNP j:

$$m(x|\, SNP_i^j) = \lambda^j(x)\exp(\gamma^j SNP_i^j), \tag{13}$$

where the function $\lambda^j(x)$ is left unspecified. Having calculated for each SNP the corresponding

coefficient, the PRS of individual i can now be estimated as

$$PRS_i = SNP_i^1 \hat{\gamma}^1 + SNP_i^2 \hat{\gamma}^2 + \ldots + SNP_i^S \hat{\gamma}^S, \tag{14}$$

where S denotes the number of the significant SNPs. Alternatively, we could estimate another version of the PRS by using our new model. Therefore,

$$m(x|SNP_i^j) = a^j \exp\left(b^j \exp(c^j * SNP_i^j)x\right). \tag{15}$$

For this specification, the PRS of individual i can be estimated as

$$PRS_i = SNP_i^1 \hat{c}^1 + SNP_i^2 \hat{c}^2 + \ldots + SNP_i^S \hat{c}^S. \tag{16}$$

Another possibility for the calculation of the PRS is to follow a statistical machine learning approach, where the calculation of the number of significant SNPs for the estimation of the PRS would be chosen based on the (heuristic) minimization of a loss function. The procedure for choosing the optimal number, S, of SNPs for the estimation of the PRS consists of three steps after dividing the sample into a training set and a test set. The first step is to rank the SNPs based on the corresponding $p-$ values obtained by employing either regression (13) or (15). Then we could consider the first ρ SNPs from that step onwards, where $\rho << 1200K$.

The (heuristically) optimal choice of the S SNPs from the set of the ρ candidate SNPs could be made by using stepwise forward selection. In summary, after having ranked the SNPs, we could construct the $\tau-$th PRS by considering the first τ SNPs, where $\tau \in \{1, \ldots, \rho\}$, starting from the first SNP (after having ranked them). Let PRS_i^τ be the PRS that is estimated by using the first τ SNPs for individual i. For the sake of exposition, we stick below with the Cox model for the estimation of the PRS as well as the estimation of the effect of the PRS on the mortality rate. Assuming that the weights have been obtained by employing the Cox regression (13), we have in

mathematical notation

$$PRS_i^{\tau} = SNP_i^1 \hat{\gamma}^{1} + SNP_i^2 \hat{\gamma}^{2} + \ldots + SNP_i^{\tau} \hat{\gamma}^{\tau}. \tag{17}$$

The main challenge here is to choose the value of l for which the PRS_i^l has the optimal predictive power. Nevertheless, there is one major difficulty with that approach. Mortality outcomes are often (heavily) randomly right censored, and thus it is not trivial how the predictive power of the PRS can be evaluated in the second stage. One solution to this problem is to write down a Cox regression for the censoring variable. In particular, let $m(x|PRS_i^{\tau})$ and $m_C(x|PRS_i^{\tau})$ be the hazard rates of the mortality variable and the censoring variable C_i, respectively, for the $i-$th individual given the value PRS_i^{τ}. Those hazard rates can be modeled as

$$m(x|PRS_i^{\tau}) = \lambda^{\tau}(x) \exp(\psi^{\tau} PRS_i^{\tau}), \tag{18}$$

$$m_C(x|PRS_i^{\tau}) = \lambda_C^{\tau}(x) \exp(\psi_C^{l} PRS_i^{\tau}). \tag{19}$$

The second step is to estimate (by keep using the training set) for each τ the equations (18) and (19). In the third step, we should use the test set observations in order to evaluate the predictive power of the PRS^{τ}. More precisely, we have from equations (18), (19) that

$$\ln \int_0^x m(\omega|PRS_i^{\tau}) = -\psi^{l} PRS_i^{\tau} + E, \tag{20}$$

and

$$\ln \int_0^x m_C(\omega|PRS_i^{\tau}) = -\psi_C^{l} PRS_i^{\tau} + E_C, \tag{21}$$

where E and E_C follow an extreme value distribution. The quantities on the left as well as right hand side (except for the error terms, of course) have been consistently estimated from the second step. Hence, for fixed τ and each (un)censored observation with PRS_i^{τ} in the test set, we can predict the corresponding (un)censored observation for PRS_i^{τ}. By repeating the same procedure for each τ and developing an appropriate loss function, the optimal value of τ, which will minimize that loss function, can be determined.

Appendix

Recall that the mortality rate is equal to

$$m(x|PRS) = a\exp(b\exp(c*PRS)x). \tag{A-1}$$

Therefore, the survival function $S(x|PRS)$ can be written as $S(x|PRS) = \exp(-M(x|PRS))$, where $M(x|PRS) := \int_0^x m(\omega|PRS)\,d\omega$. Direct calculations reveal that

$$M(x|PRS) = \frac{a}{b\exp(c*PRS)}(\exp(b\exp(c*PRS)x) - 1). \tag{A-2}$$

The latter result implies that for each $x > 0$

$$S(x|PRS) = \exp\left(-\frac{a}{b\exp(c*PRS)}(\exp(b\exp(c*PRS)x) - 1)\right). \tag{A-3}$$

Accordingly, the probability density function is expressed as

$$f(x|PRS) = a\exp(b\exp(c*PRS)x)\exp\left(-\frac{a}{b\exp(c*PRS)}(\exp(b\exp(c*PRS)x) - 1)\right). \tag{A-4}$$

After having calculated the PRS_i for the $i-$th individual, our data consist of the triplet

$$(\min(X_i, C_i), PRS_i, D_i),$$

where C_i is the censoring variable and D_i is an indicator function equal to 1 for uncensored observations (i.e. $X_i < C_i$) and 0 for censored observations (*i.e.* $X_i \geq C_i$). The individual likelihood contribution is equal to

$$D_i f(X_i|PRS_i) + (1 - D_i)S(X_i|PRS_i).$$

Next, we carry out a simple Monte Carlo experiment to investigate how the fully parametric estimator performs. Our numerical study is based on the following parameter values: $a = 0.005$, $b = 0.09$, $c = 0.07$. In addition, we assume that the variable PRS is uniformly distributed over the

interval $[-1, 1]$. We also generate a censoring variable that depends on PRS, too. In particular, we choose the following hazard rate for the censoring variable

$$m_C(x|PRS) = a\exp(b * 1.15 * \exp(c * 0.9 * PRS)x). \quad \text{(A-5)}$$

The resulting average degree of censoring is about 32% for all Monte-Carlo experiments. The simulation experiments are conducted using a series of three different sample sizes: $n = 100$, $n = 200$, and $n = 500$. The number of replicated samples for each n is equal to 100. The results for each n are summarized in the Table 4. [1]

Sample Size	$Bias(\hat{a})$	$RMSE(\hat{a})$	$Bias(\hat{b})$	$RMSE(\hat{b})$	$Bias(\hat{c})$	$RMSE(\hat{c})$
n=100	4.419 e-05	0.0028	2.055 e-03	0.0171	3.65 e-03	0.068
n=200	1.330 e-05	6.046 e-05	6.994 e-04	0.001	-3.04 e-03	0.0093
n=500	5.694 e-05	0.0033	-1.471 e-03	0.0148	-3.86 e-03	0.0308

Table 4: Bias and Mean Square Error results for the three parameter estimates.

References

Aly, M., Wiklund, F., Xu, J., Isaacs, W. B., Eklund, M., DAmato, M., Adolfsson, J., and Gro¨nberg, H. (2011), "Polygenic risk score improves prostate cancer risk prediction: results from the Stockholm-1 cohort study," *European urology*, 60, 21–28.

Belsky, D. W., Moffitt, T. E., Baker, T. B., Biddle, A. K., Evans, J. P., Harrington, H., Houts, R., Meier, M., Sugden, K., Williams, B., et al. (2013), "Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study," *JAMA psychiatry*, 70, 534–542.

[1] For the numerical implementation of the maximum likelihood estimation, we choose to replace each of the parameters a, b, and c with the following fraction

$$\frac{s_l \exp(\omega_l)}{1 + \exp(\omega_l)},$$

where $s_l > 0$ are fixed and the maximization is performed over the parameters ω_l for $l \in \{a, b, c\}$. This means that we can transform our constrained optimization problem into an unconstrained optimization problem. In our simulations, the latter generally outperforms the former in terms of bias and RMSE.

Cross-Disorder Group of the Psychiatric Genomics Consortium and others (2013), "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *The Lancet*, 381, 1371–1379.

Derks, E. M., Vorstman, J. A., Ripke, S., Kahn, R. S., Ophoff, R. A., Consortium, S. P. G., et al. (2012), "Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: a polygenic risk score analysis," *PloS one*, 7, e37852.

Dudbridge, F. (2013), "Power and predictive accuracy of polygenic risk scores," *PLoS Genet*, 9, e1003348.

Ganna, A., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., Magnusson, P. K., Pedersen, N. L., Ingelsson, E., and Tiemeier, H. (2013), "Genetic determinants of mortality. Can findings from genome-wide association studies explain variation in human mortality?" *Human genetics*, 132, 553–561.

Hamad, R. and Rehkopf, D. (2015), "Telomere Length and Health: A Genetic instrumental Variable Analysis," *Working Paper*.

Levine, M. E., Crimmins, E. M., Prescott, C. A., Phillips, D., Arpawong, T. E., and Lee, J. (2014), "A polygenic risk score associated with measures of depressive symptoms among older adults," *Biodemography and social biology*, 60, 199–211.

Sebastiani, P., Solovieff, N., DeWan, A. T., Walsh, K. M., Puca, A., Hartley, S. W., Melista, E., Andersen, S., Dworkis, D. A., Wilk, J. B., et al. (2012), "Genetic signatures of exceptional longevity in humans," *PloS one*, 7, e29848.

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., Kraft, P., Chen, R., Kallberg, H. J., Kurreeman, F. A., et al. (2012), "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis," *Nature genetics*, 44, 483–489.

Vaupel, J., Manton, K., and Stallard, E. (1979), "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, 16, 439–454.

Vink, J. M., Hottenga, J. J., Geus, E. J., Willemsen, G., Neale, M. C., Furberg, H., and Boomsma, D. I. (2014), "Polygenic risk scores for smoking: predictors for alcohol and cannabis use?" *Addiction*, 109, 1141–1151.

Walter, S., Mackenbach, J., Vok´o, Z., Lhachimi, S., Ikram, M. A., Uitterlinden, A. G., Newman, A. B., Murabito, J. M., Garcia, M. E., Gudnason, V., et al. (2012), "Genetic, physiological, and lifestyle predictors of mortality in the general population," *American Journal of Public Health*, 102, e3–e10.

Yashin, A. I., Wu, D., Arbeev, K. G., and Ukraintseva, S. V. (2012), "Polygenic effects of common single-nucleotide polymorphisms on life span: when association meets causality," *Rejuvenation research*, 15, 381–394.