Biodemographic approaches to genetic analyses of longevity in longitudinal data on aging

Konstantin G. Arbeev[1], Olivia Bagley[1], Ilya Y. Zhbannikov[1], Anatoliy I. Yashin[1]

[1]Biodemography of Aging Research Unit (BARU), Social Science Research Institute, Duke University, Durham, NC, USA

Abstract
Modern longitudinal studies often collect genetic information in addition to follow-up data on mortality or other events. Typically, individuals are genotyped at different ages, and the demographic structure of the genotyped population provides additional information about the effect of genetic variants on the event of interest (along with follow-up data on genotyped and non-genotyped individuals). We present the general genetic-demographic approach which takes such structure into account and describe results of simulation studies which illustrate that combining information on follow-up and information on ages at biospecimen collection improves power in analyses of genetic effects on mortality compared to analyses of follow-up data alone. We also illustrate the approach in application to a genome-wide association study (GWAS) of lifespan in Cardiovascular Health Study (CHS) with genetic data from the CHS Candidate Gene Association Resource. We found that groups of individuals with different values of weighted polygenic risk scores (above/below median) constructed from the top SNPs in GWAS of lifespan (with p-value threshold 0.01) differ in chances to stay free of Alzheimer's disease thus validating further exploration of these findings in analyses of larger scale genetic data.

Key words: longitudinal data; genetics of longevity; mortality; polygenic risk score; biodemography; genetic-demographic model

1. Introduction

Estimation of effects of genetic markers on time-to-event outcomes (such as mortality risk or incidence of a disease) usually involves application of appropriate survival analysis methods to a sample of genotyped individuals. Such analyses can also benefit from a "genetic-demographic" (GD) approach (Yashin et al. 1999, Yashin et al. 2000, Dato, Carotenuto, and De Benedictis 2007, Yashin, Arbeev, and Ukraintseva 2007, Arbeev et al. 2011) that takes into account the demographic structure of the genotyped population under study. Usually genetic data are collected from participants of longitudinal studies at different ages. Then the allele-/genotype-specific age structure of the population at the time of biospecimen collection (i.e., proportions of carriers of different alleles/genotypes at different ages) also conveys information about the effect of genetic variants on the event of interest, in addition to follow-up data. Indeed, if the outcome is mortality, then an individual needs to survive until the age at biospecimen collection in order to be genotyped, or he/she should be event-free at that time if the event of interest is incidence of a disease. In addition, the non-genotyped part of the study provides additional information on the events, and the non-genotyped sample consists of the mixture of carriers and non-carriers of the alleles or genotypes being studied. Hence, the use of these additional sources of information can improve power compared to the analyses of follow-up data on genotyped individuals alone (Yashin, Arbeev, and Ukraintseva 2007, Arbeev et al. 2011).

The rest of the text is organized as follows. Section 2 describes the general GD model. This presentation extends the original GD approach (Yashin et al. 1999, Yashin et al. 2000, Arbeev et al. 2011) allowing both allele-/genotype-specific survival functions and initial proportions of alleles/genotypes to depend on additional covariates. Section 3 presents results of simulation studies which illustrate that combining information on follow-up and information on ages at biospecimen collection improves power in analyses of genetic effects on mortality/morbidity risks compared to analyses of follow-up data alone. Section 4 shows the application results of the approach to data on mortality in the Cardiovascular Health Study (CHS) with genetic data from the CHS Candidate Gene Association Resource (CARe). Section 5 concludes the paper.

2. General "Genetic-Demographic" Model

2.1 Some Notations

Consider a study with N independent individuals at the baseline and let $N = N_{gen} + N_{nongen}$, where N_{gen} and N_{nongen} are the numbers of genotyped and non-genotyped individuals in the sample, respectively. Let G_i be a random variable with values g_i, $g_i = 1\ldots G$, denoting the presence of some allele or genotype in the genome of i^{th} individual. For example, it may be a binary variable coding the presence (1) or absence (0) of the minor allele at some locus, or it may be a variable counting the number of the minor alleles coded as 0, 1, and 2. For genotyped individuals, information on the genetic marker is available (i.e., the value g_i is known for i^{th} individual) but for non-genotyped individuals this value is unknown.

We assume that time-to-event information and other relevant covariates are available for both genotyped and non-genotyped individuals. Denote τ_i age at death/censoring, δ_i a censoring indicator (0 if censored, 1 if died), t_i^0 age at baseline, and X_i a (column) vector of covariates for i^{th} individual from the sample, $i = 1\ldots N$. Let t_i^{gen}, $i = 1\ldots N_{gen}$, be ages at biospecimen collection for genotyped individuals. In a general case, biospecimen collection can happen after some time since the baseline, so these two ages can be different (the simpler situation when biospecimen are collected at the baseline is a special case of the general formulas which will be mentioned below).

Denote by $\mu(t \mid G_i = g_i, X_i)$ the hazard rate for an individual with alleles/genotypes g_i (whether observed or not) and a vector of covariates X_i and let $S_{g_i}(t \mid X_i)$ be the respective survival function:

$$S_{g_i}(t \mid X_i) = P(\tau_i > t \mid G_i = g_i, X_i) = \exp\left\{-\int_0^t \mu(u \mid G_i = g_i, X_i)du\right\}. \quad (1)$$

Individuals with different alleles/genotypes g_i entering the study at age t_i^0 (and having the values of covariates X_i) have, in general, different chances to survive until the age at biospecimen collection t_i^{gen}. Denote the proportion of individuals with allele/genotype g_i who survived until the age at biospecimen collection t_i^{gen} given that they entered the study at age t_i^0 and have the values of covariates X_i as $\pi_{g_i}(t_i^{gen} \mid t_i^0, X_i) = P(G_i = g_i \mid \tau_i > t_i^{gen}, t_i^0, X_i)$.

2.2 Likelihood Function for Genotyped Subsample

For ith individual from the genotyped subsample, we observe his/her (censored) lifespan (τ_i, δ_i) and information on the genetic variant of interest (g_i), conditional on having the individual's age at baseline t_i^0, age at biospecimen collection t_i^{gen}, and a vector of covariates X_i.

The initial probabilities $P(G_i = k \mid X_i)$ can be represented, for example, using a multinomial logistic regression (it could be any other functional relationships between the covariates and the probability, see, e.g., Sections 3 and 4):

$$P(G_i = k \mid X_i) = \frac{e^{\beta_{0k} + \beta_{1k}^T X_i}}{1 + \sum_{c=1}^{G-1} e^{\beta_{0c} + \beta_{1c}^T X_i}}, \qquad (2)$$

for $k = 1 \ldots G-1$, and

$$P(G_i = G \mid X_i) = 1 - \sum_{k=1}^{G-1} P(G_i = k \mid X_i) = \frac{1}{1 + \sum_{k=1}^{G-1} e^{\beta_{0k} + \beta_{1k}^T X_i}}. \qquad (3)$$

Here β_{0k} and β_{1k} are the intercept and the column vector of allele- or genotype-specific regression parameters, respectively, and "T" denotes transposition. Here we postulated $\beta_{0G} = 0$ and $\beta_{1G} = 0$ for identifiability (Proust-Lima et al. 2014).

The term in the likelihood function that corresponds to the genotyped subsample is

$$L_{gen} = \prod_{i=1}^{N_{gen}} L_i^{gen}, \qquad (4)$$

where

$$L_i^{gen} = \pi_{g_i}(t_i^{gen} \mid t_i^0, X_i) \mu(\tau_i \mid G_i = g_i, X_i)^{\delta_i} \exp\left\{ - \int_{t_i^{gen}}^{\tau_i} \mu(t \mid G_i = g_i, X_i) dt \right\} \qquad (5)$$

and $\pi_{g_i}(t_i^{gen} \mid t_i^0, X_i) = P(G_i = g_i \mid \tau_i > t_i^{gen}, t_i^0, X_i)$ is given by

$$\pi_{g_i}(t_i^{gen} \mid t_i^0, X_i) = \frac{P(\tau_i > t_i^{gen} \mid G_i = g_i, t_i^0, X_i) P(G_i = g_i \mid t_i^0, X_i)}{\sum_{k=1}^{G} P(\tau_i > t_i^{gen} \mid G_i = k, t_i^0, X_i) P(G_i = k \mid t_i^0, X_i)} \qquad (6)$$

with

$$P(\tau_i > t_i^{gen} \mid G_i = g_i, t_i^0, X_i) = \exp\left\{ -\int_{t_i^0}^{t_i^{gen}} \mu(t \mid G_i = g_i, X_i)dt \right\} \qquad (7)$$

and

$$P(G_i = g_i \mid t_i^0, X_i) = \frac{S_{g_i}(t_i^0 \mid X_i)P(G_i = g_i \mid X_i)}{\sum_{k=1}^{G} S_k(t_i^0 \mid X_i)P(G_i = k \mid X_i)} \qquad (8)$$

where $S_{g_i}(t_i^0 \mid X_i)$ are given by (1).

2.3 Likelihood Function for Non-Genotyped Subsample

We assume that the genotyped and non-genotyped subsamples are independent and that they are representative to each other, that is, carriers/non-carriers of the respective alleles/genotypes in these subsamples have the same parameters of hazard rates (note that if this is not the case, but a functional relationship between the parameters in the genetic and non-genetic subsamples can be reasonably assumed, then this situation can also be modeled). We also assume that the initial proportions are given by (2), (3).

For *j*th individual from the non-genotyped subsample, *j* = 1… N_{nongen}, we observe his/her (censored) lifespan (τ_j, δ_j), conditional on having the individual's age at baseline t_j^0 and a vector of covariates X_j. Information on the genetic variant of interest is unknown for the non-genotyped individuals.

Therefore, the term in the likelihood function that corresponds to the non-genotyped subsample is

$$L_{nongen} = \prod_{j=1}^{N_{nongen}} L_j^{nongen}, \qquad (9)$$

where

$$L_j^{nongen} = \sum_{k=1}^{G} \mu(\tau_j \mid G_j = k, X_j)^{\delta_j} \exp\left\{ -\int_{t_j^0}^{\tau_j} \mu(t \mid G_j = k, X_j)dt \right\} P(G_j = k \mid t_j^0, X_j) \qquad (10)$$

and $P(G_j = k \mid t_j^0, X_j)$ is given by (8).

2.4 Likelihood Function for Combined Genotyped and Non-Genotyped Subsamples

Since participants of the study with and without genetic information are assumed to be independent from each other, the combined likelihood function is

$$L = L_{gen} L_{nongen}. \qquad (11)$$

An important property of the likelihood terms (4) and (9) is that they are based on the same specifications of hazard for carriers of different alleles/genotypes, and, therefore, they have the same unknown parameters. This property suggests that the joint analysis of data from genotyped and non-genotyped subsamples by maximizing the likelihood (11) will improve the accuracy of parameter estimates compared to the estimates evaluated in the analyses of data from the genotyped subsample alone (i.e., maximizing the likelihood (4)).

2.5 Special Case when Biospecimen Are Collected at Baseline

In case when biospecimen are collected at the baseline examination, we have $t_i^{gen} = t_i^0$ and the respective term L_i^{gen} in the likelihood (4) simplifies to

$$L_i^{gen} = \mu(\tau_i \mid G_i = g_i, X_i)^{\delta_i} \exp\left\{ -\int_{t_i^0}^{\tau_i} \mu(t \mid G_i = g_i, X_i) dt \right\} P(G_i = g_i \mid t_i^0, X_i), \quad (12)$$

where $P(G_i = g_i \mid t_i^0, X_i)$ is given by (8).

3. Simulation Studies

In our simulation studies, we assumed that the hazard rate for individuals with different number of alleles is $\mu(x \mid G) = \mu_0(x) e^{\beta_G G + \beta_X^T X}$, where the variable *G* counts the number of the alleles of interest, the baseline mortality $\mu_0(x)$ is the Gompertz function, i.e., $\ln \mu_0(x) = \ln a + bx$, and *X* is a vector with two covariates representing birth cohort (simulated as 1950-X_0, where X_0 is age at baseline exam, uniformly distributed over the interval [30, 60]) and sex (0 or 1, with probability 0.5). The initial distribution of genotypes in a population is assumed according to the Hardy-Weinberg equilibrium and the probability of having an allele of interest is modeled as: $\text{logit}(\pi) = \beta_0 + \beta_P^T X$. We used the following parameters: $\ln a = -9.0$, $b = 0.08$, $\beta_G = 0.4$, $\beta_X^T = (-0.005, 0.2)$, $\beta_0 = -1.0$, $\beta_P^T = (-0.01, 0.2)$ which provide realistic mortality curves similar to contemporary populations.

We generated a “general population” of 10,000,000 individuals, assigning values of *G* to individuals computed in accordance with the probabilities for respective values of covariates *X*. Then we generated lifespans for all individuals from the corresponding probability distributions (i.e., those corresponding to the hazard for individuals carrying *G* alleles and having covariates *X*, with the parameters defined above). Then we assigned a hypothetical “age at entry” into the study to each individual in the population generated as a discrete random variable uniformly distributed over the interval 30 to 60 years. We assumed that individuals were genotyped 40 years after the baseline and that the follow-up period was 60 years. We collected a sample of 5,000 individuals whose life spans exceeded their hypothetical “age at entry.” Individuals with simulated lifespans exceeding “age at entry” plus the follow-up period were considered censored at that age. We assumed that 1,250 individuals constitute the “genotyped” sample for which the values of *G* are known and the rest of the sample is non-genotyped so that these values are unknown (but still lifespan information is available). Such design resembles the structure of the Framingham Heart Study (original cohort) (Dawber, Meadors, and Moore 1951). This procedure was repeated 100 times to generate 100 datasets which were estimated using the likelihood function presented in the previous section. The results are shown in Table 1.

Table 1 about here

To illustrate the increase in power in case of combining information on follow-up and information on ages at biospecimen collection compared to analyses of follow-up data alone, we performed simulation studies with different values of effect size (parameter β_G). We simulated a scenario with a shorter follow up period (8 years) and a wider range of ages at baseline (40 to 100), with genotyping at the baseline (i.e., all 5,000 individuals are genotyped), to resemble a more common situation in contemporary longitudinal studies (such as Long Life Family Study, see, e.g., Yashin et al. 2010). All other parameters were selected as indicated above. We compare the approach which takes into account differential survival of individuals with different genotypes vs. the approach which uses only the follow-up information (with the left truncation defined as the age at baseline which ignores the fact that carriers of different numbers of alleles *G* have different chances to survive until the baseline to become the participants of the study). Fig. 1 indicates that the former approach results in substantial improvement in power compared to the latter one.

Fig. 1 about here

4. Applications

We applied the method to the Cardiovascular Health Study (CHS) data. The CHS is a population-based, longitudinal study of risk factors for the development and progression of heart disease and stroke in the Medicare-eligible older individuals aged 65+ years at enrollment (Fried et al. 1991). The main cohort of 5,201 study participants was examined annually from 1989 through 1999. In June 1993, an additional 687 African Americans were recruited using similar methods. Deaths were ascertained through surveillance and at semi-annual contacts (Fried et al. 2001). In this study, we used the CHS data provided by the database of Genotypes and Phenotypes (dbGaP), dbGaP Study Accession: phs000287.v5.p1. We focused on the subsample of whites in CHS (referred to as CHS-W below). The CHS-W sample includes data on 4,648 individuals (2,607 females, 2,041 males) aged 65-100 years at the baseline exam.

We used the Candidate Gene Association Resource (CARe) data provided by dbGaP (dbGaP Study Accession: phs000377.v5.p1) which include information on genotyping of about 50,000 single nucleotide polymorphisms (SNPs) in approximately 2,100 candidate genes and pathways for cardiovascular, inflammatory and metabolic phenotypes, done using the same customized Illumina's iSelect array (the IBC-chip) in each study of the CARe project (Keating et al. 2008, Musunuru et al. 2010).

We applied the quality control (QC) procedure (Anderson et al. 2010) to CHS-W CARe data. We removed variants with minor allele frequency < 0.01, Hardy–Weinberg equilibrium P-value < 0.00001, and genotype failure rate > 0.05. We excluded all individuals with a genotype failure rate ≥ 0.05 or a heterozygosity rate ± 3 standard deviations from the mean, and individuals with a first or second principal component (PC) score ± 8 standard deviations from the mean reference population (CEU+TSI HapMap3 individuals). PCs used in QC and in analyses were computed using R-package

GENESIS (Conomos, Miller, and Thornton 2015). The resulting sample after QC contained data on 4,183 individuals (2,360 females, 1,823 males) and 34,411 SNPs from autosomal chromosomes. We applied the model described in Section 2 to these data (using the likelihood for genotyped individuals L_{gen}). The following specification was used: $\mu(x \mid G) = \mu_0(x) e^{\beta_G G + \beta_X^T X}$, with the Gompertz baseline mortality and the vector X with two covariates representing birth cohort (date of birth minus the minimal year of birth in the study (1885) grouped in 5 year intervals (0 = 1885-1889 through 40 = 1925-1929)) and sex. The initial probabilities of having an effect allele are modeled as $\text{logit}(\pi) = \beta_0 + \beta_P^T X_P$, with X_P containing birth cohort, sex and first PC.

The results of applications of this model are shown in Fig. 2.

Fig. 2 about here

We also constructed (weighted) polygenic risk scores (PRS) using the results of GWAS of lifespan in CHS-W CARe data described above (with p-value threshold 0.01) and evaluated how participants with different values of PRS (above/below median) differ in their chances to stay free of Alzheimer's disease (defined from CHS hospital discharge data, ICD9-CM codes 331.0 and 290.1x). The results are shown in Fig. 3.

Fig. 3 about here

5. Conclusions

We presented the general genetic-demographic (GD) model that takes into account the demographic structure of the genotyped population (i.e., proportions of carriers of different alleles/genotypes at different ages) in analyses of effects of genetic markers on time-to-event outcomes (e.g., mortality). The model allows both allele-/genotype-specific survival functions and initial proportions of alleles/genotypes to depend on additional covariates thus extending the original GD approach (Yashin et al. 1999, Yashin et al. 2000, Arbeev et al. 2011). The demographic structure of the genotyped population conveys information about the effect of genetic variants on the event of interest, thus its incorporation in the analyses improves power compared to the analyses of follow-up data on genotyped individuals alone, as we illustrated in simulation studies presented here. This effect is especially noticeable in the studies with shorter follow-up, as we showed earlier using the original GD approach (Yashin et al. 2013). Application of the GD approach to GWAS of lifespan in CHS CARe data (subsample of whites) did not reveal genome-wide significant signals. Nevertheless, we found that groups of individuals with different values of weighted PRS (above/below median) constructed from the top SNPs in GWAS of lifespan (with p-value threshold 0.01) differ in chances to stay free of Alzheimer's disease (the effect is observed in both females and males) thus validating further exploration of these findings in analyses of larger scale genetic data.

Acknowledgements

This work was partly supported by the National Institute on Aging of the National Institutes of Health under Award Numbers P01AG043352, P30AG034424, and R01AG046860. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

In this work we used the Cardiovascular Health Study (CHS) data provided by the database of Genotypes and Phenotypes (dbGaP), dbGaP Study Accession: phs000287.v5.p1. For genetic analyses, we used the Candidate Gene Association Resource (CARe) data (dbGaP Study Accession: phs000377.v5.p1).

CHS contract acknowledgment: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01-HC-85079, N01-HC-85080, N01-HC-85081, N01-HC-85082, N01-HC-85083, N01-HC-85084, N01-HC-85085, N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, and N01-HC-85239; grant numbers U01 HL080295 and U01 HL130014 from the National Heart, Lung, and Blood Institute, and R01 AG-023629 from the National Institute on Aging, with additional contribution from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at https://chs-nhlbi.org/pi. This manuscript was not prepared in collaboration with CHS investigators and does not necessarily reflect the opinions or views of CHS, or the NHLBI.

CHS CARe acknowledgement: Support for the genotyping through the CARe Study was provided by NHLBI Contract N01-HC-65226.

References

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. 2010. "Data quality control in genetic case-control association studies." *Nature Protocols* no. 5 (9):1564-1573. doi: 10.1038/nprot.2010.116.

Arbeev, K. G., S. V. Ukraintseva, L. S. Arbeeva, I. Akushevich, A. M. Kulminski, and A. I. Yashin. 2011. "Evaluation of genotype-specific survival using joint analysis of genetic and non-genetic subsamples of longitudinal data." *Biogerontology* no. 12 (2):157-66. doi: 10.1007/s10522-010-9316-1.

Conomos, M. P., M. B. Miller, and T. A. Thornton. 2015. "Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness." *Genetic Epidemiology* no. 39 (4):276-293. doi: 10.1002/gepi.21896.

Dato, S., L. Carotenuto, and G. De Benedictis. 2007. "Genes and longevity: a genetic-demographic approach reveals sex- and age-specific gene effects not shown by the case-control approach (APOE and HSP70.1 loci)." *Biogerontology* no. 8 (1):31-41. doi: 10.1007/s10522-006-9030-1.

Dawber, T. R., G. F. Meadors, and F. E. Moore. 1951. "Epidemiological approaches to heart disease: The Framingham Study." *American Journal of Public Health* no. 41 (3):279-286.

Fried, L. P., N. O. Borhani, P. Enright, C. D. Furberg, J. M. Gardin, R. A. Kronmal, L. H. Kuller, T. A. Manolio, M. B. Mittelmark, A. Newman, and et al. 1991. "The Cardiovascular Health Study: design and rationale." *Ann Epidemiol* no. 1 (3):263-76.

Fried, L. P., C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke, and M. A. McBurnie. 2001. "Frailty in older adults: evidence for a phenotype." *J Gerontol A Biol Sci Med Sci* no. 56 (3):M146-56.

Keating, B. J., S. Tischfield, S. S. Murray, T. Bhangale, T. S. Price, J. T. Glessner, L. Galver, J. C. Barrett, S. F. Grant, D. N. Farlow, H. R. Chandrupatla, M. Hansen,

S. Ajmal, G. J. Papanicolaou, Y. Guo, M. Li, S. Derohannessian, P. I. de Bakker, S. D. Bailey, A. Montpetit, A. C. Edmondson, K. Taylor, X. Gai, S. S. Wang, M. Fornage, T. Shaikh, L. Groop, M. Boehnke, A. S. Hall, A. T. Hattersley, E. Frackelton, N. Patterson, C. W. Chiang, C. E. Kim, R. R. Fabsitz, W. Ouwehand, A. L. Price, P. Munroe, M. Caulfield, T. Drake, E. Boerwinkle, D. Reich, A. S. Whitehead, T. P. Cappola, N. J. Samani, A. J. Lusis, E. Schadt, J. G. Wilson, W. Koenig, M. I. McCarthy, S. Kathiresan, S. B. Gabriel, H. Hakonarson, S. S. Anand, M. Reilly, J. C. Engert, D. A. Nickerson, D. J. Rader, J. N. Hirschhorn, and G. A. Fitzgerald. 2008. "Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies." *PLoS One* no. 3 (10):e3583. doi: 10.1371/journal.pone.0003583.

Musunuru, Kiran, Guillaume Lettre, Taylor Young, Deborah N. Farlow, James P. Pirruccello, Kenechi G. Ejebe, Brendan J. Keating, Qiong Yang, Ming-Huei Chen, Nina Lapchyk, Andrew Crenshaw, Liuda Ziaugra, Anthony Rachupka, Emelia J. Benjamin, L. Adrienne Cupples, Myriam Fornage, Ervin R. Fox, Susan R. Heckbert, Joel N. Hirschhorn, Christopher Newton-Cheh, Marcia M. Nizzari, Dina N. Paltoo, George J. Papanicolaou, Sanjay R. Patel, Bruce M. Psaty, Daniel J. Rader, Susan Redline, Stephen S. Rich, Jerome I. Rotter, Herman A. Taylor, Jr., Russell P. Tracy, Ramachandran S. Vasan, James G. Wilson, Sekar Kathiresan, Richard R. Fabsitz, Eric Boerwinkle, Stacey B. Gabriel, and Nhlbi Candidate Gene Assoc. 2010. "Candidate Gene Association Resource (CARe): Design, Methods, and Proof of Concept." *Circulation. Cardiovascular Genetics* no. 3 (3):267-275.

Proust-Lima, C., M. Sene, J. M. Taylor, and H. Jacqmin-Gadda. 2014. "Joint latent class models for longitudinal and time-to-event data: a review." *Statistical Methods in Medical Research* no. 23 (1):74-90. doi: 10.1177/0962280212445839.

Yashin, A. I., K. G. Arbeev, and S. V. Ukraintseva. 2007. "The accuracy of statistical estimates in genetic studies of aging can be significantly improved." *Biogerontology* no. 8 (3):243-255. doi: 10.1007/s10522-006-9072-4|ISSN 1389-5729.

Yashin, A. I., G. De Benedictis, J. W. Vaupel, Q. Tan, K. F. Andreev, I. A. Iachine, M. Bonafe, M. DeLuca, S. Valensin, L. Carotenuto, and C. Franceschi. 1999. "Genes, demography, and life span: The contribution of demographic data in genetic studies on aging and longevity." *American Journal of Human Genetics* no. 65 (4):1178-1193. doi: 10.1086/302572.

Yashin, A. I., G. De Benedictis, J. W. Vaupel, Q. Tan, K. F. Andreev, I. A. Iachine, M. Bonafe, S. Valensin, M. De Luca, L. Carotenuto, and C. Franceschi. 2000. "Genes and longevity: Lessons from studies of centenarians." *Journals of Gerontology Series A Biological Sciences and Medical Sciences* no. 55 (7):B319-B328.

Yashin, Anatoli I., Konstantin G. Arbeev, Alexander Kulminski, Ingrid Borecki, Kaare Christensen, Michael Barmada, Evan Hadley, Winifred Rossi, Joseph H. Lee, Rong Cheng, and Irma T. Elo. 2010. ""Predicting" parental longevity from offspring endophenotypes: Data from the Long Life Family Study (LLFS)." *Mechanisms of Ageing and Development* no. 131 (3):215-222. doi: 10.1016/j.mad.2010.02.001.

Yashin, Anatoliy I, Konstantin G. Arbeev, Deqing Wu, Liubov S. Arbeeva, Alexander M Kulminski, Igor Akushevich, Irina Culminskaya, Eric Stallard, and Svetlana Ukraintseva. 2013. "How the quality of GWAS of human lifespan and health span can be improved." *Frontiers in Genetics* no. 4:article 125. doi: 10.3389/fgene.2013.00125.

Figures:

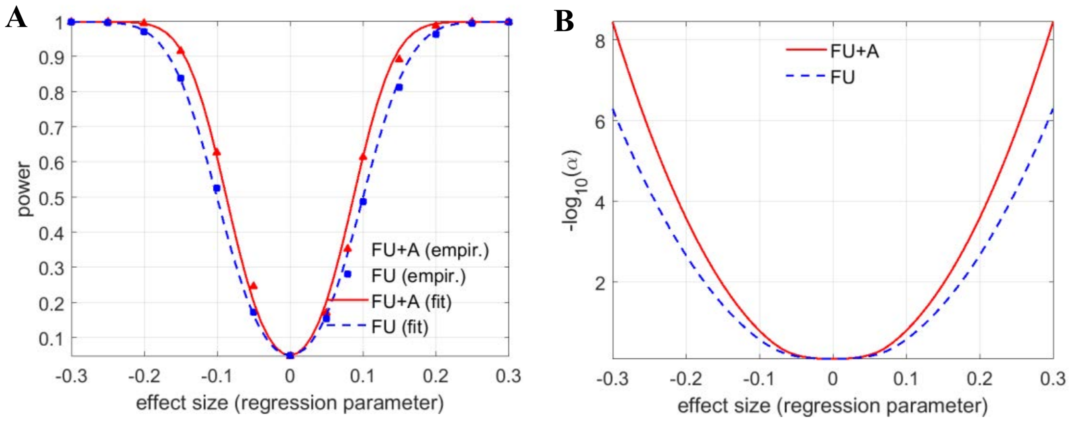

Figure 1: (A) Power in two approaches (using data on follow-up only, “FU”, and data on follow-up and ages at biospecimen collection, “FU+A”) for different effect sizes (i.e., values of regression parameter γ) and fixed $\alpha = 0.05$. Markers (“empir.”) denote values from simulations and lines (“fit”) correspond to power curves of a one-sample Z-test of the mean (with standard deviations producing the best fit to simulated values in two approaches: 0.051 and 0.044, respectively). **(B)** Level of the test (shown as $-\log_{10}(\alpha)$ for better visibility) that yields power *w*=0.8, as a function of the effect size in both approaches (the curves are calculated using the values of standard deviations mentioned above).

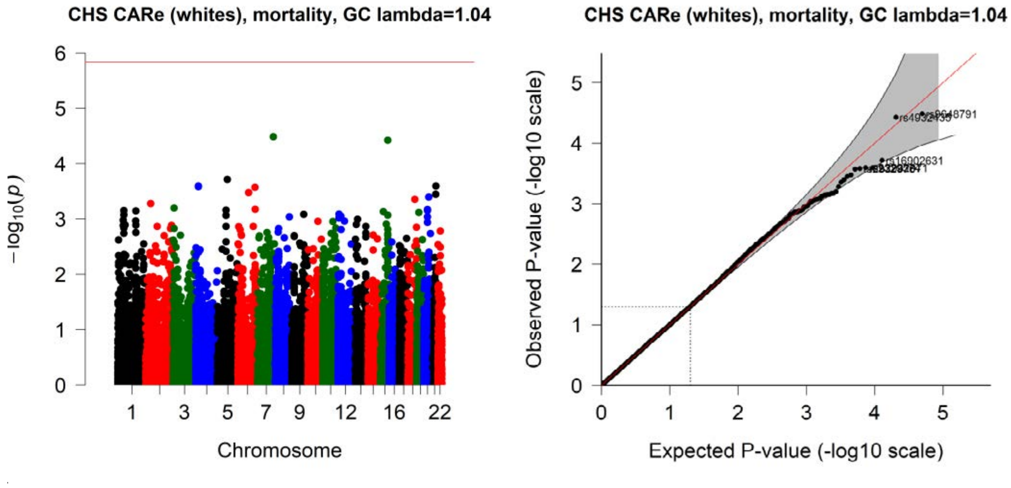

Figure 2: Results of GWAS of lifespan in CHS CARe data, subsample of whites (model adjusted for sex, birth cohort, PC1)

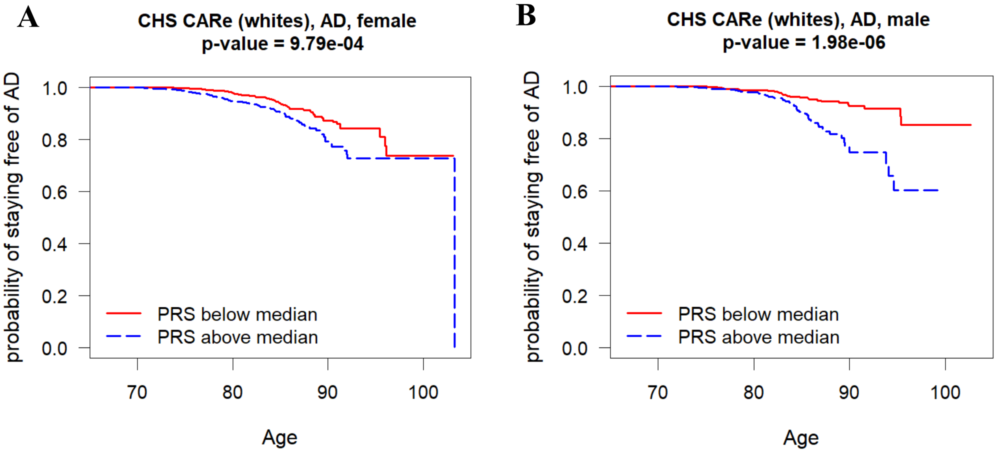

Figure 3: Kaplan-Meier curves for probabilities of staying free of Alzheimer's disease for individuals from CHS-W CARe data with different values of polygenic risk score (PRS): **A)** females; **B)** males. PRS were computed from GWAS of lifespan in CHS-W CARe, with p-value threshold 0.01.

Tables:

Table 1: Results of simulation studies

	$\ln a$	$b \times 10$	β_G	$\beta_X(1) \times 100$	$\beta_X(2)$	β_0	$\beta_P(1) \times 10$	$\beta_P(2)$
Mean	-9.047	0.804	0.410	-0.442	0.197	-0.908	-0.129	0.188
St. Dev.	0.148	0.015	0.045	0.199	0.029	0.231	0.089	0.090
Min	-9.451	0.768	0.297	-0.914	0.112	-1.489	-0.317	-0.054
Max	-8.723	0.844	0.543	0.003	0.263	-0.351	0.102	0.404
True Values	***-9.0***	***0.8***	***0.4***	***-0.5***	***0.2***	***-1.0***	***-0.1***	***0.2***