Biodemographic Approaches Can Improve Power of Genetic Analyses of Longitudinal Data on Aging, Health, and Longevity

Konstantin G. Arbeev[1*], Liubov S. Arbeeva[1], Igor Akushevich[1], Alexander M. Kulminski[1], Deqing Wu[1], Svetlana V. Ukraintseva[1,2], Irina V. Culminskaya[1], Anatoliy I. Yashin[1,2]

[1]Social Science Research Institute, Duke University, Durham, NC 27705, USA
[2]Duke Cancer Institute, Duke University, Durham, NC 27710, USA

[*] **Corresponding Author:** Dr. Konstantin G. Arbeev, Duke University, Social Science Research Institute, 2024 W. Main St., Room A102F, Box 90420, Durham, NC 27705, USA, ka29@duke.edu

Abstract

We discuss different approaches to work with rich data available in modern longitudinal studies of aging, health, and longevity that started collecting genetic information in addition to follow-up data on events and longitudinal measurements of biomarkers. Such methods provide a possibility to improve the power of genetic analyses by joint analysis of data for genotyped and non-genotyped sub-samples of the study. We describe results of simulation studies in the longitudinal genetic-demographic model illustrating that inclusion of information on ages at biospecimen collection in addition to follow-up data improves power in analyses of genetic effects on mortality/morbidity risks. We present simulation studies in the genetic stochastic process model illustrating the increase in power in joint analyses of genotyped and non-genotyped participants compared to analyses of non-genotyped participants alone in different scenarios testing relevant biologically-based hypotheses. We illustrate applications of these approaches to analyses of genetic data in the Framingham Cohorts.

1. Introduction

The modern era of revolutionary advances in genetics provides great opportunities and challenges for the field of biodemography and the need to integrate the principles of genetics and genomics into biodemography is apparent so that this field would continue to be on the forefront of the demographic analyses [1, 2]. The importance of "genetic biodemography" will continue to grow in the coming years because many studies that collected data on biomarkers will include (or already have included) genetic information. The ongoing incorporation of genetic information into longitudinal studies is considered potentially "the most revolutionary element of the addition of biological data in large-scale surveys" [3] and such studies will "increasingly provide analyses of the interactions of genetic, biological, social, economic, and demographic characteristics" [4].

To get the full advantage of such rich data, a special attention should be paid to the analytic approaches to work with this diverse information. Consider, for example, the situation when the research interest is in the evaluation of the genetic effect on some time-to-event outcome, e.g., risk of death or onset of a disease. Comparison of the age patterns of incidence or mortality rates for carriers of different alleles/genotypes can help understand the role of genetic

factors in survival or development of aging-associated diseases. Traditional methods to estimate the effect of genetic markers in such cases can be enhanced if we complement them with the demographic approach taking into account the demographic structure of the population under study. Specifically, when genetic data are included in longitudinal studies of aging, we have several relevant sources of information for analyses of genetic influence on lifespan (or onset of diseases), in addition to genetic data themselves.

First, it is follow-up data on the outcome of interest (e.g., mortality). Second, usually genetic data are collected in longitudinal studies from participants at different ages. Therefore, this provides information on age structure of the population at the time of biospecimen collection. Along with follow-up data, such population age structure also contains information about the effect of genetic variants on lifespan and the full potential of the data is underused when this information is ignored in analyses, especially when genotyping is performed at advanced ages with noticeable attrition due to mortality. Indeed, in order to be genotyped, an individual has to survive until the age at biospecimen collection. Hence, if the proportion of carriers of some genetic variant increases with age (here we mean the age at biospecimen collection) then this variant should favor longevity. This implies that we can associate genetic variants with lifespan even without the follow-up data using the "gene frequency" method [5, 6]. We can expect therefore that if we use both follow-up data and data on population age structure then this would provide us with more accurate estimates of parameters and additional power compared to the use of follow-up data alone. Such data can be analyzed jointly using appropriate methods [7, 8].

The third source of information in longitudinal studies of aging that is relevant for genetic analyses stems from the history of incorporation of genetic information into such studies. While in some modern longitudinal studies the genetic data can be collected at the baseline, it is a common situation that many older long-established longitudinal studies started before the genetic data collection began. Hence, in such studies genetic data are available only for a sub-sample of participants of the longitudinal study (i.e., for those who survived until the time of biospecimen collection). It is also possible that genetic data were collected only for a sub-sample of participants due to, for example, budgetary restrictions. However, in both such cases information on the outcome of interest (e.g., follow-up on mortality) can be available for all (genotyped and non-genotyped) participants of the longitudinal study. This information should not be neglected in genetic analyses because it provides an additional reserve for increasing power and improving the accuracy of the estimates. Indeed, the group of non-genotyped individuals is a mixture of carriers/non-carriers of the same alleles/genotypes collected in the genetic data and a similar functional form of mortality rate can be assumed for the entire sample. Therefore, this information can be appropriately combined in the likelihood function with information for genotyped individuals [8].

Incorporation of genetic information in the studies that collect longitudinal measurements of biomarkers along with follow-up data opens new perspectives for analyses of genetic influence on aging, health and longevity. Participants of a longitudinal study for whom genetic information was not collected but other outcomes (longitudinal measurements of biomarkers, follow-up data, and possibly some other relevant covariates) are still available, provide an additional source to increase the accuracy and power in analyses of genetic effects on longitudinal and time-to-event outcomes. The approach to jointly analyze longitudinal measurements of biomarkers and time-to-event outcomes for genotyped and non-genotyped

participants of longitudinal studies has been developed recently within the framework of the stochastic process model (SPM) of aging [9, 10]. Such a model, named the "genetic stochastic process model," or the "genetic SPM," is especially relevant in the context of biodemographic research. The particular advantage of the genetic SPM for biodemographic applications is that it is based on biological theory and this model incorporates several essential mechanisms of aging-related changes in organisms and it allows for evaluating genetic effects on such characteristics and their influence on mortality or onset of a disease. Such "hidden components" of aging-related changes incorporated into this model include: adaptive capacity, resistance to stresses, physiological norm, and effects of allostatic adaptation. As known from the literature, all these variables play important roles in the processes of aging. Therefore, their inclusion in the model is crucial for better understanding of regulatory mechanisms driving observed aging-related changes in physiological variables and their influence on risks of death or getting a disease, as well as for evaluating the genetic component in such processes. However, relevant variables associated with such "hidden components of aging" are typically not directly measured in longitudinal data and, hence, they cannot be directly estimated from the data using, for example, joint models. The genetic SPM thus provides a useful approach to work with such "hidden components of aging" indirectly. Importantly, it also provides an additional possibility to improve the power of genetic analyses by joint analysis of data for genotyped and non-genotyped sub-samples of the study [9].

The rest of the paper is organized as follows. Section 2 presents results of simulation studies in the longitudinal genetic-demographic model [8] illustrating that inclusion of information on ages at biospecimen collection in addition to follow-up data improves power in analyses of genetic effects on mortality or morbidity risks (see also [11]). Section 3 presents modified version of the genetic SPM [9] that includes the dependence of the model's components on the vector of observed (time-independent) covariates available at baseline and describes simulation studies illustrating the increase in power in joint analyses of genotyped and non-genotyped participants of a longitudinal study compared to analyses of non-genotyped participants alone in different scenarios to test relevant biologically-based hypotheses. Section 4 discusses some applications of the approaches to real data. Section 5 discusses the results and possible generalizations of the approaches.

2. Simulation Studies in Longitudinal Genetic-Demographic Model

The longitudinal genetic-demographic model (or the genetic-demographic model for longitudinal data) is described in Arbeev et al. [8]. The full model combines three sources of information in the likelihood function: 1) follow-up data on survival (or, generally, on some time-to-event) for genotyped individuals; 2) (cross-sectional) information on ages at biospecimen collection for genotyped individuals; and 3) follow-up data on survival for non-genotyped individuals. In the simulation study presented in this section, we utilize only the first two sources. Of course, follow-up information for non-genotyped individuals provides an additional reserve for improving the power of genetic analyses but this simulation study illustrates that, even for the studies where genetic data are collected for all participants, the use of information on ages at biospecimen collection still makes a difference for the power of genetic analyses.

Let x_k^0, $k = 1\ldots K$, be the ages at baseline (entry to the study) of individuals from the genotyped subsample of the data and let x_{m,x_k^0}, $m = 1\ldots M_k$, be their ages at the time of biospecimen collection. Denote by $N(x_{m,x_k^0}) = N_1(x_{m,x_k^0}) + N_0(x_{m,x_k^0})$ the number of individuals in the genotyped subsample who were aged x_{m,x_k^0} at the time of biospecimen collection and aged x_k^0 at baseline. Here $N_g(x_{m,x_k^0})$ are the numbers of non-carriers ($g = 0$) and carriers ($g = 1$) of some allele/genotype. Let τ denote the life span (it may be censored). Denote by $\mu(x \mid G = g)$ the hazard rate for carriers/non-carriers and by $\pi(x_{m,x_k^0} \mid x_k^0) = P(G = 1 \mid \tau > x_{m,x_k^0}, x_k^0)$ the proportion of carriers at age x_{m,x_k^0} given that the individuals were aged x_k^0 at baseline. Denote by $S_g(x) = P(\tau > x \mid G = g)$ the survival functions for carriers/non-carriers and by $P_1 = P(G = 1)$ the initial proportion (at birth) of carriers of the allele/genotype in a population, which is assumed here to be the same for different birth cohorts represented in the study. The total (population) survival function is then $S(x) = P_1 S_1(x) + (1 - P_1) S_0(x)$. Conditional survival functions for the individuals aged x_k^0 at the baseline are $S_g(x \mid x_k^0) = P(\tau > x \mid G = g, x_k^0)$. The hazard rates for carriers/non-carriers can be of any parametric form, e.g., the Gompertz curves as in our simulations presented below. The proportions $\pi(x_{m,x_k^0} \mid x_k^0)$ are:

$$\pi(x_{m,x_k^0} \mid x_k^0) = \frac{P(G = 1 \mid x_k^0) S_1(x_{m,x_k^0} \mid x_k^0)}{P(G = 1 \mid x_k^0) S_1(x_{m,x_k^0} \mid x_k^0) + (1 - P(G = 1 \mid x_k^0)) S_0(x_{m,x_k^0} \mid x_k^0)}, \quad (1)$$

where $P(G = 1 \mid x_k^0) = P_1 S_1(x_k^0) / S(x_k^0)$.

The likelihood function of the data on the ages at biospecimen collection (L_A) and the likelihood function of the follow-up data (L_{FU}) are [8]:

$$L_A \sim \prod_{k=1}^{K} \prod_{m=1}^{M_k} \pi(x_{m,x_k^0} \mid x_k^0)^{N_1(x_{m,x_k^0})} (1 - \pi(x_{m,x_k^0} \mid x_k^0))^{N_0(x_{m,x_k^0})} \quad (2)$$

and

$$L_{FU} \sim \prod_{k=1}^{K} \prod_{m=1}^{M_k} \prod_{g=0}^{1} \prod_{i=1}^{N_g(x_{m,x_k^0})} \mu(\tau_i \mid G = g)^{\delta_i} S_g(\tau_i \mid x_{m,x_k^0}), \quad (3)$$

where δ_i is a censoring indicator. The total likelihood function of the data relevant for genetic analyses of the genotyped subsample is the product of these two likelihood functions:

$$L_{FU+A} \sim L_{FU} L_A. \quad (4)$$

In our simulation studies we compared two methods of estimating parameters of the allele- or genotype-specific hazard rates: 1) the method that uses only follow-up data, i.e., the likelihood function L_{FU} (3); and 2) the method that uses both data on the ages at biospecimen collection and follow-up data, i.e., the likelihood function L_{FU+A} (4).

We assumed that carriers and non-carriers of some hypothetical allele in a population have mortality rates $\mu(x \mid G) = \mu_0(x)e^{\gamma G}$, where the variable G denotes carriers (G = 1) or non-carriers (G = 0), the baseline mortality $\mu_0(x)$ is the Gompertz function, i.e., $\ln \mu_0(x) = \ln a + bx$, with $\ln a = -10.0$ and $b = 0.09$, and the proportion of carriers at birth P_1 = 0.25. We varied the parameter γ from -0.5 to 0.5 with the interval 0.05 to simulate scenarios with different effect sizes.

We generated a "general population" of 10,000,000 individuals assigning the genetic status (i.e., variable G) to individuals in accordance with the initial proportion P_1. Then we generated life spans for all individuals from the respective probability distributions (i.e., those corresponding to the hazard $\mu_0(x)e^{\gamma G}$ for carriers and $\mu_0(x)$ for non-carriers, with the parameters defined above). Then we assigned the hypothetical "age at entry" into the study to each individual in the population generated as a discrete random variable uniformly distributed over the interval 40 to 100 years. We assumed that individuals were genotyped at the baseline, i.e., their age at biospecimen collection coincides with age at entry. We collected a sample of 4,500 individuals whose life spans exceeded their hypothetical "age at entry." We considered two scenarios: a short follow-up period (6 years) and a long follow-up period (60 years). Individuals with simulated life spans exceeding "age at entry" plus respective follow-up period were considered censored at that age in the respective scenario (note that in the scenario with a long follow-up period almost all individuals experienced the event whereas in the scenario with a short follow-up period a substantial proportion of individuals is censored). This procedure was repeated 1,000 times (in each scenario with different γ and follow-up period) to generate 1,000 datasets which were estimated using the likelihoods (3) and (4).

Fig. 1 (A) illustrates the empirical power (i.e., the proportion of datasets in which the null hypothesis H0: $\gamma = 0$ was rejected at $\alpha = 0.05$) in the scenario with a short follow-up and for different effect sizes (i.e., values of the parameter γ). We also fitted these empirical values with the power curves of a one-sample Z-test of the mean and found the values of the standard deviations that produced the best fit to the empirical power curves for each method (0.059 for "FU+A" (4) and 0.088 for "FU" (3)). Fig. 1 (B) shows the 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 methods (the curves were calculated using the abovementioned values of standard deviations). Fig. 1 (C) and Fig. 1 (D) display similar quantities for the scenario with a long follow-up period.

Fig. 1 (A) and Fig. 1 (B) illustrate that, in case of a short follow-up period, the use of information on ages at biospecimen collection in addition to follow-up data gives a substantial increase in power compared to the traditional approach that uses the follow-up data alone. For example, Fig. 1 (B) shows that for the effect size $\gamma = 0.3$ p-value reduces approximately from 10^{-2} to 10^{-5} and for the effect size equal to $\gamma = 0.4$ p-value drops approximately from 10^{-4} to 10^{-9}. This means that many genetic variants which would not reach the genome-wide significance in genome-wide association studies (GWAS) using the traditional approach analyzing the follow-up data alone could become highly significant if the data on ages at biospecimen collection were also used. Fig. 1 (C) and Fig. 1 (D) reveal that this effect diminishes for a long follow-up period. In the case of a long follow-up period information from this long follow-up makes a more substantial contribution compared to information hidden in the distributions of ages at biospecimen collection. Conversely, in the case of a short follow-up period, distributions of the

ages at biospecimen collection play a more important role in differentiating the allele- or genotype-specific survival patterns compared to the follow-up data (in the case of a substantial proportion of censored individuals, as in our simulations).

Our simulations thus illustrate that the additional use of information on ages at biospecimen collection may have important implications for GWAS of longevity or onset of diseases in cases with short follow-up periods (which are the majority of data currently available).

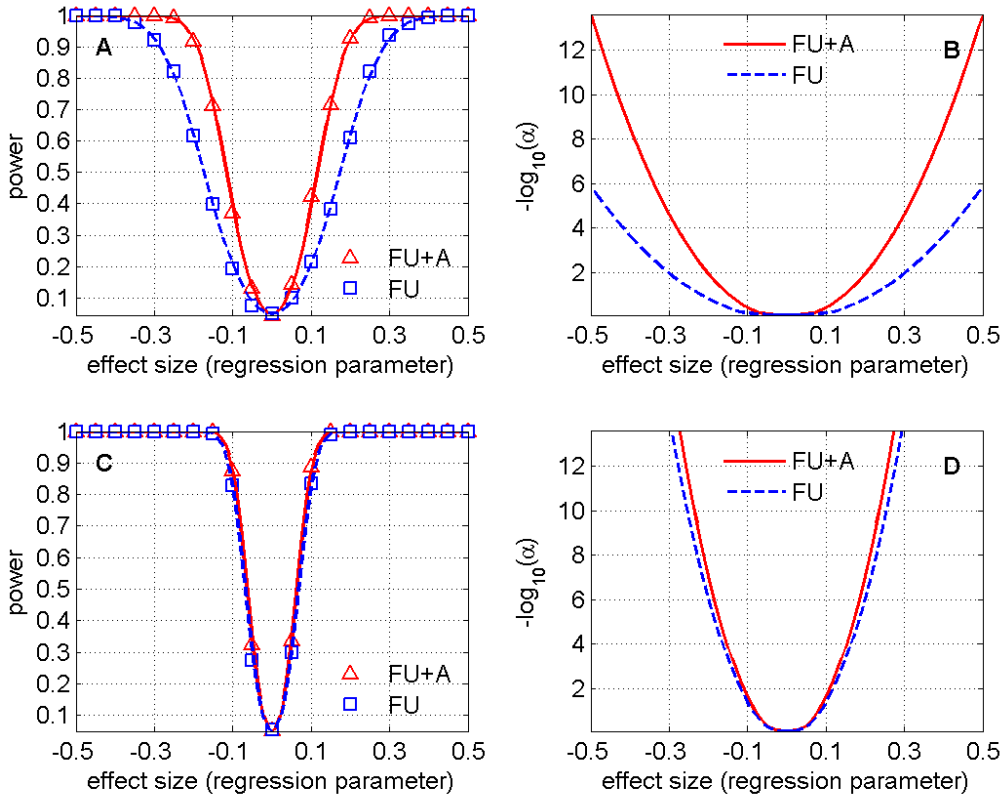

Figure 1: Simulation studies in longitudinal genetic-demographic model: **(A)** Power in two methods (with follow-up only, "FU", and follow-up and ages at biospecimen collection, "FU+A") for different effect sizes (i.e., values of the regression parameter γ) and $\alpha = 0.05$ in the scenario with a short follow-up period (6 years). The lines denote the fit of the empirical curves by the power curves of a one-sample Z-test of the mean (the standard deviations that produced the best fit are 0.059 for "FU+A" and 0.088 for "FU"). **(B)** The 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 methods (the curves are calculated using the abovementioned values of standard deviations) in the scenario with a short follow-up period (6 years). **(C)** is same as (**A**) but for a long follow-up period (60 years). The standard deviations that produced the best fit are 0.032 for "FU+A" and 0.035 for "FU." **(D)** is same as (**B**) but for a long follow-up period (60 years).

3. Simulation Studies in Genetic Stochastic Process Model

The genetic stochastic process model was developed in Arbeev et al. [9]. Here we present its version modified to include the dependence of the model's components on the vector of observed (time-independent) covariates available at baseline and describe simulation studies to illustrate the increase in power in joint analyses of genotyped and non-genotyped participants of a longitudinal study compared to analyses of only non-genotyped participants in different scenarios to test relevant biologically-based hypotheses.

Let g, $g = 1 \ldots G$, denote the presence of allele/genotype g in the genome of an individual. We can specify the probabilities of having this allele/genotype, p_g, conditional on some vector of time-independent covariates X. One possibility, for example, is to specify this probability using a multinomial logistic regression:

$$p_g = \frac{e^{\beta_{0g} + \beta_{1g}^T X}}{1 + \sum_{c=1}^{G-1} e^{\beta_{0c} + \beta_{1c}^T X}}, \tag{5}$$

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

$$p_G = \frac{1}{1 + \sum_{g=1}^{G-1} e^{\beta_{0g} + \beta_{1g}^T X}}. \tag{6}$$

Here "T" denotes transposition (we will use column vectors if not stated otherwise).

Let Y_t (t is age) be the stochastic process representing age dynamics of an M-dimensional vector of biomarkers in carriers of allele/genotype g with the following stochastic differential equation:

$$dY_t = a(t, g, X)(Y_t - f_1(t, g, X))dt + B(t, g, X)dW_t, \tag{7}$$

with initial condition Y_{t_0}. Here W_t is an M-dimensional vector Wiener process independent of the vector of initial values Y_{t_0} which represents external (and unobserved) disturbances affecting the trajectory of biomarkers. The strength of external disturbances is characterized by the $M \times M$ matrix of diffusion coefficients $B(t, g, X)$. The vector-function $f_1(t, g, X)$ (having the same dimension as Y_t) introduces the notion of allostasis into the model representing the age trajectories of biomarkers that organisms are forced to follow by the process of allostatic adaptation (see detailed description of the meaning of different components of the stochastic process model in Arbeev et al. [9]). The negative feedback coefficient in equation (7), the $M \times M$ matrix $a(t, g, X)$, describes the adaptive (homeostatic) capacity in an aging organism. The elements of this matrix correspond to the rate of adaptive response to any deviation of trajectories Y_t from the trajectories $f_1(t, g, X)$.

The hazard rates for carriers of allele/genotype g conditional on the vector of biomarkers Y_t and the vector of observed covariates X are given as:

$$\mu(t \mid Y_t, g, X) = \mu_0(t, g, X) + (Y_t - f_0(t, g, X))^T Q(t, g, X)(Y_t - f_0(t, g, X)). \tag{8}$$

Here $\mu_0(t,g,X)$ is the baseline hazard for carriers of allele/genotype g characterizing the risk that would remain if the vector Y_t followed the trajectory $f_0(t,g,X)$, and $Q(t,g,X)$ is a non-negative-definite symmetric $M \times M$ matrix. The M-dimensional vector-function $f_0(t,g,X)$ introduces the concept of age-dependent physiological norm into the model and it corresponds to the values of biomarkers which minimize the risk at respective age for carriers of allele/genotype g. The matrix $Q(t,g,X)$ in the quadratic hazard term can be associated with the decline in resistance to stresses with age, as discussed in Yashin et al. [12, 13] and Arbeev et al. [14].

The likelihood function for the model (5)-(8) is a straightforward modification of the likelihood for the original model in Arbeev et al. [9] and is not presented here. Note that the likelihood function contains the parts for the genotyped and non-genotyped sub-samples and that both parts contain the same parameters of the model. Hence, the use of available information from the non-genotyped participants (i.e., the longitudinal measurements of biomarkers and time-to-event data) provides an opportunity for increasing the power compared to analyses based on the genotyped sample alone. The advantage of the genetic stochastic process model is that it has different components which represent specific biological concepts and aging-related mechanisms for which the respective parameters have clear biological interpretations. Dependence of the model's components on variable g allows for formulating and testing different hypotheses on the presence of genetic effect of the alleles/genotypes on respective aging-related characteristics (such as stress resistance, adaptive capacity, age-dependent physiological norms, etc.). Below we present the results of simulation study that compares the power for testing of several such hypotheses in two approaches: 1) using only information from the genotyped participants; and 2) in joint analyses of the genotyped and non-genotyped individuals.

We used the following specifications of the model's components in simulations: 1) Gompertz baseline hazards: $\ln \mu_0(t,g,X) = \ln a_{\mu_0}^g + b_{\mu_0}^g t + \beta_X^g X$, where g = 1, 2 for carriers and non-carriers of a hypothetical allele (genotype), $X = c - c_0$, c is year of birth (cohort), c_0 = 1890, in simulation #2 (see Table 1) and $\ln \mu_0(t,g,X) = \ln a_{\mu_0}^g + b_{\mu_0}^g t$ in the other simulations; 2) linear functions for the multipliers in the quadratic hazard: $Q(t,g,X) = a_Q^g + b_Q^g t$; 3) linear functions for the mean allostatic trajectories: $f_1(t,g,X) = a_{f_1}^g + b_{f_1}^g t$; 4) linear functions for physiological norms: $f_0(t,g,X) = a_{f_0}^g + b_{f_0}^g t$; 5) linear functions for the negative feedback coefficient in (7) representing the adaptive capacity of an organism: $a(t,g,X) = a_Y^g + b_Y^g t$, with $a_Y^g \le 0$ and $b_Y^g \ge 0$, 6) constant diffusion coefficients: $B(t,g,X) = \sigma_1^g$; 7) normally distributed initial values of the process Y_t with means $f_1(t_0^j,g,X)$ (where t_0^j is age at the first exam for j^{th} individual) and standard deviations σ_0^g; and 8) initial probability of carrying the allele/genotype (p_1) is independent of covariates X. The values of the respective parameters were chosen to provide realistic samples resembling real data on mortality in the Framingham Original cohort data [15] and with longitudinal dynamics Y_t mimicking pulse pressure. See Table 1 summarizing parameters in different simulation studies.

We performed six simulation studies for testing different biological hypotheses on genetic effects on aging-related characteristics (see columns "**Null Hypothesis**" and

"**Interpretation of Null Hypothesis**" in Table 2). In each scenario, we simulated 100 datasets with data on age at death/censoring and the longitudinal dynamics of Y_t for 2,500 individuals followed up for 60 years with ages at baseline uniformly distributed over the interval 30 to 60 years and with 30 biennial exams measuring Y_t. Year of birth c for simulation #2 was defined as 1950 minus age at baseline. We assumed that 500 individuals have been genotyped and genetic data were not available for the rest of the sample. Power was estimated as the proportion of datasets in which the respective null hypothesis was rejected at the 0.05 significance level by the likelihood ratio test (see Table 2). For these purposes, we estimated the original (unrestricted) models and the restricted models that assume that respective parameters (highlighted in Table 1) are equal for carriers and non-carriers (simulation #2 assumes the restriction $b_X^g = 0$). Column "**Gen. Only**" in Table 2 corresponds to the likelihood that used only information from the genotyped participants and column "**Gen. + Non-Gen.**" displays the power for the likelihood with joint analyses of the genotyped and non-genotyped individuals. The table shows that joint analysis of the genotyped and non-genotyped individuals allows for a substantial increase in the power compared to analyses based on information from the genotyped participants alone thus making possible to reveal genetic effects on aging-related characteristics which would remain non-significant in analyses of the genotyped sample.

Table 1. Simulation studies in genetic stochastic process model: Parameters used to generate data (parameters used to define respective null hypotheses to be tested in each simulation are ***highlighted***)

Simulation	G	Baseline Hazard ($\mu_0(t,g,X)$)			Quadr. Hazard ($Q(t,g,X)$)		Adaptive Capacity ($a(t,g,X)$)		Mean Allostatic Trajectory ($f_1(t,g,X)$)		Physiological Norm ($f_0(t,g,X)$)		Other Parameters		
		$\ln a^g_{\mu_0}$	$b^g_{\mu_0}$	β^g_X	a^g_Q	b^g_Q	a^g_Y	b^g_Y	$a^g_{f_1}$	$b^g_{f_1}$	$a^g_{f_0}$	$b^g_{f_0}$	σ^g_0	σ^g_1	p_1
1	1	***-9.0***	***0.080***		0.5	0.1	-0.25	1.0	45.0	0.20	45.0	0.1	5.0	4.0	0.25
	2	***-8.5***	***0.082***		0.3	0.1	-0.20	1.0	50.0	0.25	40.0	0.1	5.0	4.0	
2	1	-9.0	0.080	***-0.014***	0.5	0.1	-0.25	1.0	45.0	0.20	45.0	0.1	5.0	4.0	0.25
	2	-8.5	0.082	***-0.014***	0.3	0.1	-0.20	1.0	50.0	0.25	40.0	0.1	5.0	4.0	
3	1	-9.0	0.080		***0.5***	***0.1***	-0.25	1.0	45.0	0.20	45.0	0.1	5.0	4.0	0.25
	2	-8.5	0.082		***0.5***	***0.4***	-0.20	1.0	50.0	0.25	40.0	0.1	5.0	4.0	
4	1	-9.0	0.080		0.5	0.1	***-0.22***	***1.0***	45.0	0.20	45.0	0.1	5.0	4.0	0.25
	2	-8.5	0.082		0.3	0.1	***-0.20***	***1.0***	50.0	0.25	40.0	0.1	5.0	4.0	
5	1	-9.0	0.080		0.5	0.1	-0.25	1.0	***45.0***	***0.20***	45.0	0.1	5.0	4.0	0.25
	2	-8.5	0.082		0.3	0.1	-0.20	1.0	***46.0***	***0.20***	40.0	0.1	5.0	4.0	
6	1	-9.0	0.080		0.5	0.1	-0.25	1.0	45.0	0.20	***50.0***	***0.1***	5.0	4.0	0.25
	2	-8.5	0.082		0.3	0.1	-0.20	1.0	50.0	0.25	***40.0***	***0.1***	5.0	4.0	

Notes:

1) Some parameters are rescaled for better visibility in the table: a^g_Q is multiplied by 10^4; b^g_Q is multiplied by 10^5; b^g_Y is multiplied by 10^3.

Table 2. Simulation studies in genetic stochastic process model: Power (for $\alpha = 0.05$ and effect sizes defined by respective parameters from Table 1) in case of the likelihood estimating only data on genotyped individuals (column **"Gen. Only"**) and data on both genotyped and non-genotyped individuals (column **"Gen. + Non-Gen."**)

Simulation	**Null Hypothesis**	**Interpretation of Null Hypothesis**	**Power**	
			Gen. Only	**Gen. + Non-Gen.**
1	$\mu_0(t,g,X) = \mu_0(t,X)$	No genetic effect on baseline hazard	0.42	0.85
2	$\mu_0(t,g,X) = \mu_0(t,g)$	No cohort changes in baseline hazard	0.25	0.89
3	$Q(t,g,X) = Q(t,X)$	No genetic effect on stress resistance	0.41	0.90
4	$a(t,g,X) = a(t,X)$	No genetic effect on adaptive capacity	0.40	0.82
5	$f_1(t,g,X) = f_1(t,X)$	No genetic effect on mean allostatic trajectory	0.71	0.89
6	$f_0(t,g,X) = f_0(t,X)$	No genetic effect on physiological norm	0.44	0.91

4. Applications

Applications of the genetic-demographic model and the genetic SPM have been performed in our recent publications [8, 10, 16, 17]. These applications allowed making important insights on the genetics of aging and longevity and the genetic determinants of aging-related mechanisms.

Here we discuss one application of the genetic SPM related to estimating genetics of stress resistance from longitudinal data. Typically, longitudinal data on aging in humans contain limited (if any at all) information on longitudinal dynamics of biomarkers which can be associated with stress resistance and would allow for investigating the genetic component of decline in stress resistance with age. The genetic SPM allows evaluating this important aging-related component indirectly from the estimates of the U-shaped mortality risk as a function of observed covariates. Eq. (8) in the one-dimensional case is represented by a quadratic function:

$$\mu(t \mid Y_t, g, X) = \mu_0(t, g, X) + Q(t, g, X)(Y_t - f_0(t, g, X))^2. \qquad (9)$$

The value of $Q(t,g,X)$ in (9) can be associated with stress resistance because it regulates the width of the U-shaped risk function (as a function of the risk factor Y). If the value of this coefficient is small then the U-shape is wide and the risk function is less sensitive to small deviations of the risk factor Y_t from the norm ($f_0(t,g,X)$). This can be associated with better stress resistance. If the value of this coefficient is large then the U-shape is narrow and the risk function is sensitive to small deviations of Y_t from the norm which corresponds to worse stress resistance. With age, the width of the U-shape can change. For example, if it narrows with age (i.e., $Q(t,g,X)$ is an increasing function of t) then this can be linked to the phenomenon of the aging-related decline in stress resistance (see more discussion on the topic in [14]). The genetic SPM allows estimating this component ($Q(t,g,X)$) as a function of allele/genotype g, i.e., one can evaluate whether carriers and non-carriers of some specific allele/genotype differ in their stress resistance and/or the dynamics (decline) of stress resistance with age.

Fig. 2 illustrates changes in stress resistance with age for carriers and non-carriers of the APOE e4 allele in participants of the Framingham Original Cohort showing the quadratic component of the conditional hazards considered as functions of serum cholesterol (see detailed description of computations and the data in [10]). This figure shows that stress resistance (i.e., the width of the U-shaped mortality risk) among carriers of the APOE e4 allele declines faster with age than among non-carriers of this allele (and also that males tend to have worse stress resistance than females, see also [14]).

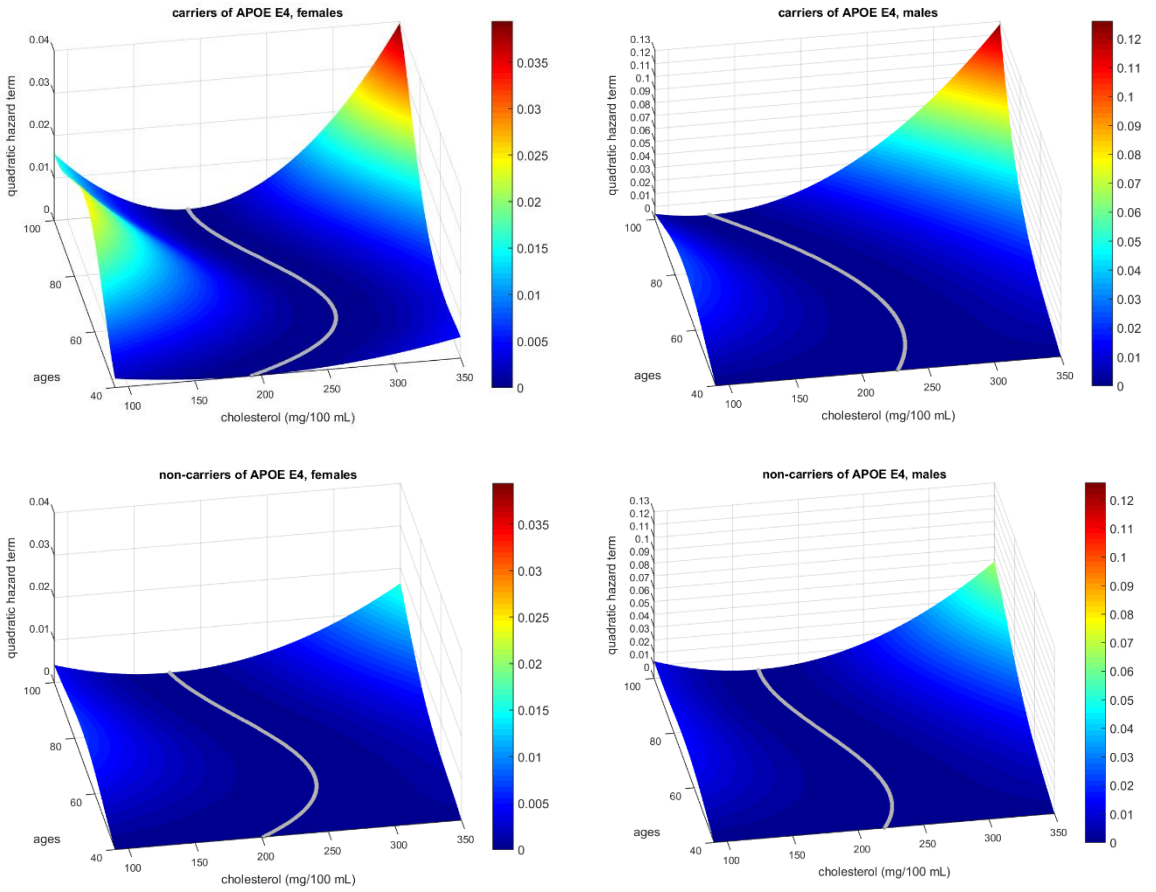

Figure 2: Application of genetic SPM to data on mortality and longitudinal dynamics of cholesterol in carriers and non-carriers of the APOE e4 allele in the Framingham Original Cohort: Changes in the U-shapes of mortality risks with age

5. Discussion

In this paper we presented different approaches that can be applied to work with rich data available in the modern longitudinal studies of aging, health, and longevity that started collecting genetic information in addition to follow-up data on events and longitudinal measurements of biomarkers.

The longitudinal genetic-demographic model described in Section 2 and in [8] provides the method to enhance genetic analyses of time-to-event outcomes from longitudinal data combining several sources of information: follow-up data on the outcome of interest (e.g., mortality) for genotyped individuals, information on age structure of the population at the time of biospecimen collection, and follow-up data on respective events for non-genotyped participants. Such joint analysis of genotyped and non-genotyped individuals can result in substantial improvements in the power and accuracy of estimates compared to analyses of genotyped subsample alone if the proportion of non-genotyped participants is large. Such situations when genetic information cannot be collected for all

participants of longitudinal studies are not uncommon. They can arise because of several reasons: 1) the longitudinal study may have started at some time before genotyping was added to the study design so that some initially participating individuals dropped out of the study (i.e., died or were lost to follow-up) by the time of genetic data collection; 2) budget constraints prohibit obtaining genetic information for the entire sample; 3) some participants refuse to provide samples for genetic analyses. Nevertheless, even in the case when genotyped individuals constitute a majority of the sample or the entire sample, application of such an approach is still beneficial in terms of the accuracy and power because it takes into account the population structure at the time of biospecimen collection which has additional information on genetic effects on the risk of death complementing the follow-up data [11]. We should still note here that, clearly, any statistical model is just an approximation of the reality and the use of even the most advanced models does not undermine the need to collect large-scale genetic data in longitudinal studies. We note also that the genetic-demographic model presented in Section 2 and in Arbeev et al. [8] uses parametric specifications of allele- or genotype-specific survival functions. More flexible specifications such as semiparametric and non-parametric models or methods that correct for unobserved heterogeneity effects can be formulated [5].

The genetic stochastic process model presented in Section 3 adds a new dimension to genetic biodemographic analyses combining information on longitudinal measurements of biomarkers available for participants of a longitudinal study in addition to follow-up data and genetic information. Such joint analyses of different sources of information collected in both genotyped and non-genotyped individuals allow for a more efficient use of the research potential of longitudinal data which otherwise remains underused if only genotyped individuals or only subsets of available information (e.g., only follow-up data on genotyped individuals) are involved in analyses. Similar to the longitudinal genetic-demographic model presented in Section 2, benefits of combining data on genotyped and non-genotyped individuals in the genetic SPM come from the presence of common parameters describing respective characteristics of the model for genotyped and non-genotyped subsamples of the data. This takes into account that the non-genotyped subsample is a mixture of carriers of the same alleles or genotypes represented in the genotyped subsample and applies the ideas of heterogeneity analyses [18]. When the non-genotyped subsample is substantially larger than the genotyped subsample then these joint analyses can lead to a noticeable increase in the power of statistical estimates of genetic parameters compared to estimates based only on information from the genotyped subsample. This approach is applicable not only to genetic data but to any discrete time-independent variable which is observed only for a subsample of individuals of a longitudinal study.

The genetic stochastic process model enhances biodemographic analyses allowing for evaluating hidden components of aging (such as age-specific physiological norms, allostasis and allostatic load, decline in adaptive capacity and stress resistance with age) that are typically not directly measured in longitudinal data and, hence, can be estimated only indirectly. Apparently, different components and mechanisms characterizing the same process of aging should be mutually dependent and work in concert. Therefore, unification of such concepts in a comprehensive model of aging is an important step forward to the

development of a systemic methodology in aging research. As the original stochastic process model, the genetic SPM allows working with several mechanisms of aging-related changes under the overarching framework of one statistical model. In addition, the genetic SPM evaluates genetic effects on such mechanisms thus providing deeper insights on genetic determinants of the processes of aging affecting mortality and morbidity risks. It permits addressing new questions in biodemographic analyses concerning genetic influence on the aging-related changes in humans, which cannot be studied using conventional approaches, for example, the joint models or standard demographic methods. Simulations in Section 3 provide several examples of hypotheses that can be tested using the genetic SPM and illustrate the differences power resulting from the addition of information on non-genotyped individuals to the analyses.

Several practical considerations should be mentioned about applications of the genetic SPM to real data. As any parametric model, the genetic SPM relies on the description of its components as specific parametric functions. Although the basic components of the model (such as the quadratic shape of the hazard, physiological norm, average allostatic trajectory, negative feedback coefficient) are all based on the solid biological theories that justify their presence in the model, the specific parametric forms of these components are unknown and may be hard to justify biologically. Moreover, these components generally cannot be empirically evaluated from the real data to guess their parametric form because they model hidden components of aging process not directly associated to any measurable variables in the data (one exception might be the baseline hazard rate which, with some degree of relevance, can be assumed to have the same shape as the hazard rate in the total population, e.g., Gompertz, Weibull, gamma-Gompertz, or gamma-Weibull baseline rates can be chosen depending on the application). Therefore, it is advisable to perform sensitivity analyses with different parametric specifications of the components of the model, e.g., linear, quadratic, or higher order polynomial functions, and select the best fitting model using formal criteria such as the likelihood ratio test for nested models or the Akaike Information Criterion for non-nested models.

The specific type of genetic influence on the hidden components of aging is not known *a priori*. Thus, versions of the model with different types of genetic influences should be tested in applications. For example, dominant, recessive, or additive form of action of the minor allele on respective characteristics can be investigated. Similarly, joint analyses of two or more genetic markers might be of interest in applications. The genetic SPM can be straightforwardly extended to work with multiple genetic markers. However, this results in a larger number of parameters and a smaller number of individuals in different groups that can reduce the reliability of estimates.

Computational burden should always be taken into account in practical implementations of statistical methods especially in large-scale problems involving studies with large sample sizes and/or extensive amounts of genetic data. For example, genome-wide association studies (GWAS) data are collected in different longitudinal studies that can contain millions of single nucleotide polymorphisms (SNPs) for thousands of participants (see the dbGaP website, http://www.ncbi.nlm.nih.gov/gap?db=gap). For such data, computational burden of the parameter estimation procedure in the genetic SPM

suggests that its routine application to each SNP in the dataset is not feasible for modern computers, especially in high dimensional cases. At the present time, a more relevant application of this model is to work with a much smaller set of SNPs pre-selected according to some criterion [19]. The likelihood estimation procedure in the longitudinal genetic-demographic model is considerably faster and, therefore, it is suitable for large-scale applications. Our experience with the version of the model by Arbeev et al. [8] indicates that estimation of GWAS data on thousands of individuals and more than a hundred thousand SNPs can be performed in a reasonable time. Nevertheless, both the genetic SPM and longitudinal genetic-demographic models can be used in studies with candidate genes or SNPs to investigate their connections with mortality risk and risks of diseases and to evaluate genetic contribution into hidden components of aging that affect these risks.

Several further generalizations of the methods to evaluate genetic influence on hidden components of aging can be considered. As discussed in Yashin et al. [20], ignoring hidden heterogeneity in a population due to the presence of latent subpopulations defined by some unobserved characteristics can lead to erroneous conclusions concerning biological regularities of aging-related processes estimated by the stochastic process model. The same, of course, is true for the genetic SPM. Therefore, the generalization of the genetic SPM to include latent classes can be useful for sensitivity analyses to test the presence of hidden heterogeneity that can affect the results of the genetic SPM.

Another direction for possible extension of the genetic SPM concerns "individualization" of longitudinal trajectories. In its present form, all individuals in the model have the same ("population") parameters of the adaptive capacity and the allostatic trajectory. Respective parameters of these components can be assumed as random variables or realizations of some stochastic process to describe individual patterns of adaptive capacity and the allostatic load. Although in this case such additional random effects and the "original" random process (i.e., the Wiener process W_t in the equation for the dynamics of longitudinal biomarker Y_t (7)) may "compete" for the same correlation structure in the longitudinal data, so that the feasibility of such an approach needs careful investigation. See also relevant discussion on the use of complicated random effects structures vs. the use of stochastic processes in the joint models literature [21, 22].

Investigation of genetic effects on hidden components of aging and their relation to risks of death and onset of diseases can also be performed in the framework of extended versions of the stochastic process model aimed at analyses of dependent competing risks [23, 24], and data on individual health histories and mortality [25] that analyze longitudinal data collected using different observational plans [26]. Such analyzes would allow addressing many new problems that cannot be investigated using standard approaches. For example, one can reveal the role of genetic factors in competing risks of death without traditional assumption on independent risks for different causes of death, investigate how genes affect hidden mechanisms of aging manifested in the longitudinal dynamics of physiological variables, and explore their relation to these dependent competing risks. The introduction of jumping components describing health states in the model allows for

comprehensive analyses of genetic effects on both fast changes in health status and slower changes in physiological state of an organism associated with aging processes, and their effects on mortality. This can help in uncovering pre-disease physiological pathways and differences in respective aging-related characteristics among carriers of different alleles or genotypes.

Acknowledgements

This work was partly supported by the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG030612, R01AG046860, P01AG043352, and P30AG034424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

References:

1. Carey, J.R., *Biodemography: Research prospects and directions.* Demographic Research, 2008. **19**: p. 1749-1757.
2. Wachter, K.W., *Biodemography comes of age.* Demographic Research, 2008. **19**: p. 1501-1512.
3. Suzman, R., *Prologue: research on the demography and economics of aging.* Demography, 2010. **47**(Supplement): p. S1-S4.
4. Crimmins, E., J.K. Kim, and S. Vasunilashorn, *Biodemography: new approaches to understanding trends and differences in population health and mortality.* Demography, 2010. **47**(Supplement): p. S41-S64.
5. Yashin, A.I., et al., *Genes, demography, and life span: The contribution of demographic data in genetic studies on aging and longevity.* American Journal of Human Genetics, 1999. **65**(4): p. 1178-1193.
6. Yashin, A.I., et al., *Genes and longevity: Lessons from studies of centenarians.* Journals of Gerontology Series A Biological Sciences and Medical Sciences, 2000. **55**(7): p. B319-B328.
7. Yashin, A.I., K.G. Arbeev, and S.V. Ukraintseva, *The accuracy of statistical estimates in genetic studies of aging can be significantly improved.* Biogerontology, 2007. **8**(3): p. 243-255.
8. Arbeev, K.G., et al., *Evaluation of genotype-specific survival using joint analysis of genetic and non-genetic subsamples of longitudinal data.* Biogerontology, 2011. **12**(2): p. 157-66.
9. Arbeev, K.G., et al., *Genetic model for longitudinal studies of aging, health, and longevity and its potential application to incomplete data.* Journal of Theoretical Biology, 2009. **258**(1): p. 103-111.

10. Arbeev, K.G., et al., *Effect of the APOE Polymorphism and Age Trajectories of Physiological Variables on Mortality: Application of Genetic Stochastic Process Model of Aging.* Scientifica, 2012. **2012**: p. Article ID 568628.
11. Yashin, A.I., et al., *How the quality of GWAS of human lifespan and health span can be improved.* Frontiers in Genetics, 2013. **4**: p. article 125.
12. Yashin, A.I., et al., *Stochastic model for analysis of longitudinal data on aging and mortality.* Mathematical Biosciences, 2007. **208**(2): p. 538-551.
13. Yashin, A.I., et al., *The quadratic hazard model for analyzing longitudinal data on aging, health, and the life span.* Physics of Life Reviews, 2012. **9**(2): p. 177-188.
14. Arbeev, K.G., et al., *Age trajectories of physiological indices in relation to healthy life course.* Mechanisms of Ageing and Development, 2011. **132**(3): p. 93-102.
15. Dawber, T.R., G.F. Meadors, and F.E. Moore, *Epidemiological approaches to heart disease: The Framingham Study.* American Journal of Public Health, 1951. **41**(3): p. 279-286.
16. Yashin, A.I., et al., *How Genes Modulate Patterns of Aging-Related Changes on the Way to 100: Lessons from Biodemographic Analyses of Longitudinal Data* in *2014 Living to 100 Monograph. SOA Monograph M–LI14–1*. 2014, Society of Actuaries: Schaumburg, IL.
17. Yashin, A.I., et al., *Genes and their role in aging.* Contingencies, 2015. **2015**(Jan.-Feb.): p. 26-33.
18. Vaupel, J.W. and A.I. Yashin, *Heterogeneity's ruses: some surprising effects of selection on population dynamics.* American Statistician, 1985. **39**(3): p. 176-185.
19. Yashin, A.I., et al., *How lifespan associated genes modulate aging changes: lessons from analysis of longitudinal data.* Frontiers in Genetics, 2013. **4**: p. article 3.
20. Yashin, A.I., et al., *Model of hidden heterogeneity in longitudinal data.* Theoretical Population Biology, 2008. **73**(1): p. 1-10.
21. Tsiatis, A.A. and M. Davidian, *Joint modeling of longitudinal and time-to-event data: An overview.* Statistica Sinica, 2004. **14**(3): p. 809-834.
22. Rizopoulos, D., *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. 2012, Boca Raton, FL: Chapman and Hall/CRC.
23. Akushevich, I., et al., *Theory of Individual Health Histories and Dependent Competing Risks.* JSM Proceedings, Section on Risk Analysis, 2011: p. 5385-5399.
24. Yashin, A.I., K.G. Manton, and E. Stallard, *Dependent competing risks: A stochastic process model.* Journal of Mathematical Biology, 1986. **24**(2): p. 119-140.
25. Yashin, A.I., et al., *Joint analysis of health histories, physiological states, and survival.* Mathematical Population Studies, 2011. **18**(4): p. 207-233.
26. Yashin, A.I., et al., *New Approach for Analyzing Longitudinal Data on Health, Physiological State, and Survival Collected Using Different Observational Plans.* JSM Proceedings, Section on Government Statistics, 2011: p. 5336-5350.