# GENETICS OF AGING, HEALTH, DISABILITY AND LONGEVITY: A STATISTICAL MODELING PERSPECTIVE

Anatoli I. Yashin, Svetlana V. Ukraintseva, Konstantin G. Arbeev,
Hai Huang, Edward Hanson
Duke University, Center for Demographic Studies, USA

**Introduction.** Traditional methods of studying the role of genetic factors in aging and longevity can be divided on two main categories. The first group of methods does not use genetic information. It uses the data on values of selected phenotypic trait (e.g. life span) measured for related individuals (e.g. twins, or other relatives). These methods assume that trait can be represented as a sum of independent genetic and environmental components. Then the component of trait's variance associated with genetic term, characterizing heritability of a trait, is evaluated. This group of methods is typical of Quantitative Genetics. It was developed when genetic data were scarce or unavailable. Although the model is extremely simplified, it often allows for detection of the presence of the genetic influence on phenotypic trait. The second group of methods actively uses genetic information. In this paper we discuss the second group of methods paying particular attention to the development of new ideas and approaches, which help better understand the nature of genetic influence on aging and aging related traits. In particular we show how demographic and epidemiological information, as well as data collected in longitudinal surveys can be used in the analysis of genetic data. We will also discuss the results of application of these approaches to data analysis.

**New Approaches to the Analysis of Genetic Data.** We illustrate the main ideas of the new approaches using the survival model for population with two genotypes. However, all procedures can easily be extended to the cases with three and more genotypes. Let $S_i(x)$, $i$=0,1 be survival functions for two genotypes representing the genetic structure of some hypothetical population and let

$$S(x)=\sum_{i=0}^{1} p_i S_i(x)$$

be a marginal survival function for an arbitrary individual in the population, where $p_i$ are initial frequencies of respective genotypes. We assume that the forces of mortality $\mu_i(x)$, $i=0,1$ for respective genotypes follow the gamma-Gompertz model, i.e.

$$\mu_i(x)=\frac{a_i e^{b_i x}}{1+s_i^2 \frac{a_i}{b_i}\left(e^{b_i x}-1\right)},\ i=0,1,$$

where $a_i, b_i$ and $s_i$ are parameters. Vaupel et al. (1979), Yashin et al. (1994), Thatcher et al. (1998) have shown that this model fits demographic mortality data better than the traditional Gompertz curve.

### The idea of the Gene Frequency (GF)-method

Let $T$ be the year of data collection, $x$ the age variable, and $N_i(x,T\text{-}x)$, $i$=0,1 the number of $x$-year-old individuals, born in the year $T$-$x$, carrying the $i$-th genotype observed in a cross-sectional study. Here $x = 0,1,\dots, X$, where $X$ is the highest age represented in the study. The cross-sectional sample represents several cohorts of individuals born in different years, and the argument $T$-$x$ merely emphasizes the fact that the counted individuals belong to the cohort born in year $T$-$x$.

Let the number "0" is associated with the candidate genotype. The number "1" is associated with any other non-"0" genotype. The method can easily be extended to the case of a population with more genotypes. However, the benefits of such an extension should always be weighed against the loss of power in the estimation procedure because of an increase in the number of parameters to be estimated.

The simplest way to assess the effect of genes on longevity using these data is to aggregate the sample into 2 age groups. The "control," or "younger," group contains all individuals under the age of 100 years. These individuals are of two genotypes "0" and *"1;"* with corresponding numbers of individuals in each group, $N_{iY}$, $i$=0,1. The "centenarian" group contains individuals aged 100 years and over. The numbers of respective genotypes in this group are $N_{iC}$ , *i=0,1*. Here the indices *"Y"* and "*C"* stand for the control (younger) and the centenarian group, respectively. The empirical estimates of relative frequencies $\hat{\pi}_{ij}$ , $i$=0,1*; j=Y,C* in these two groups are

$$\hat{\pi}_{ij} = \frac{N_{ij}}{N_{0j} + N_{1j}} \quad , i=0,1; j=Y,C$$

These estimates can be used to test the null hypothesis that the frequencies of a given genotype are the same in both age groups, against the alternative that they are not. This is, in essence, the GF-method, as it had been discussed in many publications (Takata et al. 1987, Schächter et al.1993, De Benedictis et al. 1997). This "model-free" method does not use additional information. It turns out that the use of demographic information about a population under study may substantially improve our understanding of the role of genes in aging and survival.

**Demographic Data: How Are They Helpful?** Let us consider the cohort study first. The benefits of using demographic information about marginal survival in the population results from the possibility of using the representation for the marginal survival function *S(x)* (which can be taken from the cohort demographic life tables) as a discrete mixture of respective survival functions for genotypes. Let us assume that we are dealing with two genotypes "0" and "1" as described above. Let $S_i(x)$ and $\pi_i(x)$, *i=0,1*, be the survival functions and frequencies of respective genotypes, and let $\pi_0(0)=p_0=p$ and $\pi_1(0)=p_1=1-p$ be the initial frequencies for the *"0"* and *"1"* genotypes, respectively. For simplicity we will use notation $\pi(x) = \pi_0(x)$ for the frequency of the *"0"* genotype at age *x*, and $1-\pi(x)= \pi_1(x)$ for the respective frequency of genotype *"1"* at age *x*. Then the relationships between *S(x),* $S_i(x), i=0,1,$ and $\pi(x)$ are (Vaupel and Yashin 1985):

$$S(x)=pS_0(x)+(1-p)S_1(x), \tag{1}$$

and

$$\pi(x)=pS_0(x)/(pS_0(x)+(1-p)S_1(x)). \tag{2}$$

When functions *S(x)* and $\pi(x)$ are known exactly for the age interval *[0,X]*, where *X* is the maximum age in the study, the initial gene frequency $p=\pi(0)$ for the "*0*" genotype is also known, and the survival functions for the two genotypes can be calculated from (1) and (2) by

$$S_i(x)= \pi_i(x)S(x)/p_i, \quad i=0,1. \tag{3}$$

After taking the logarithm of (3) and differentiating with respect to *x,* we get

$$\mu_i(x) = \bar{\mu}(x) - \frac{\dot{\pi}_i(x)}{\pi_i(x)}. \tag{4}$$

Thus, when the trajectories for genotype frequencies are known exactly, the addition of demographic information in the form of the marginal survival function solves the problem of determining the genetic influence on survival. One can easily show that this result does not depend on the number of genotypes observed in the study. In reality, proportions of genotypes are not known exactly. The substitution of empirical estimates of $\pi_i(x)$ and $p_i$ into (4) may create problems, because the trajectories for the estimates of survival functions for genotypes may become non-monotone (in which case respective estimates of mortality rates can have negative values). So in this case statistical methods are needed to estimate survival characteristics of genotypes.

**Merging Demographic Information with Cross-Sectional Data.** Genetic data for humans are usually collected in a cross-sectional study in some year $T$ . If the proportions of genotypes were known exactly, the use of demographic information could solve the problem of genetic influence on survival in exactly the same way as for the cohort data. Unfortunately, the proportions $\pi(x,T\text{-}x)$, of genotype *"0"* at age $x$, are not known exactly. As in the case of the GF-method we assume that $\pi(0,T\text{-}x)=p$ for all cohorts born in year $T\text{-}x$, $x=0,1,2\ldots X$. Often the numbers $N_i(x,T\text{-}x)$, $i=0,1$ are known starting with age $x^*>0$. In this case one has to assume that $\pi(x^*,T\text{-}x)=p^*$, i.e. the gene frequencies at age $x^*$ are the same for all birth cohorts. For cross-sectional data, equations (1) and (2) can therefore be rewritten as follows:

$$\tilde{S}(x) = p\tilde{S}_0(x) + (1-p)\tilde{S}_1(x)\,, \tag{5}$$

$$\tilde{\pi}(x) = p\tilde{S}_0(x)/(p\tilde{S}_0(x)+(1-p)\tilde{S}_1(x))\,. \tag{6}$$

The survival function,

$$\tilde{S}(x) = \exp\left(-\int_0^x \tilde{\mu}(u)du\right),$$

in the left hand part of (6) can be taken from a cross-sectional demographic life table for the year $T$ for the respective population, and, hence, it is a known function of $x$. Note that function $\tilde{S}(x)$ characterizes survival in the synthetic (artificial) cohort with mortality rate $\tilde{\mu}(x) = \mu(x,T-x)$ . The proportion of $x$ year old individuals carrying candidate genes in the synthetic cohort is $\tilde{\pi}(x) = \pi(x,T-x)$ . Respective survival functions for individuals carrying genotypes "0" and "1" are $\tilde{S}_i(x) = \exp\left(-\int_0^x \tilde{\mu}_i(u)du\right)$, with $\tilde{\mu}_i(x) = \mu_i(x,T-x)$ , $i=0,1$, respectively. Empirical estimates, $\hat{\pi}\ (x)$, can be calculated for each age $x=0,1\ldots X$ from the numbers $N_i(x,T\text{-}x)$, $i=0,1$, as

$$\hat{\pi}\ (x)= N_0(x,T\text{-}x)/(N_0(x,T\text{-}x)+ N_1(x,T\text{-}x)).$$

Thus, we assume that the data were obtained from binomial sampling with probability of success $\tilde{\pi}(x) = \pi(x,T-x)$ . As in the cohort case, functions $\hat{S}_i(x) = \hat{\pi}_i(x)\tilde{S}(x)/\hat{p}_i$ calculated from the analogs of (3), with $\pi_i(x)$ replaced by $\hat{\pi}_i\ (x)$, and $S(x)$ replaced by $\tilde{S}(x)$, do not necessarily decline monotonically, and, hence, strictly speaking, they cannot be considered the estimates of $\tilde{S}_i(x)$, $i=1,2$ (i.e. of survival functions for genotypes). That is why statistical methods for estimating survival functions $\widetilde{S}_i(x)$, *i=0,1* are needed. In this paper, we show that methods based on the maximum likelihood procedure can successfully be used in the joint analysis of genetic and demographic data to solve this problem.

**The Likelihood Function of Genetic Data.** The following likelihood function of genetic data is the basis for an estimation procedure in all methods discussed below:

$$L = \prod_{x=x^*}^{X} \pi(x,T-x)^{N_0(x,T-x)}(1-\pi(x,T-x))^{N_1(x,T-x)} \tag{7}$$

Here $N_0(x,T\text{-}x)$ and $N_1(x,T\text{-}x)$, $x=x^*,x^*+1,\ldots X,$ are the number of individuals with and without the genotype "0", respectively, aged $x$ at time $T$. Since we identify $\pi(x,T\text{-}x)$ with $\pi(x)$, which satisfies (D.6), the likelihood function becomes

$$L = \prod_{x=x^*}^{X}\left(\frac{p\widetilde{S}_0(x)}{p\widetilde{S}_0(x)+(1-p)\widetilde{S}_1(x)}\right)^{N_0(x,T-x)}\left(\frac{(1-p)\widetilde{S}_1(x)}{p\widetilde{S}_0(x)+(1-p)\widetilde{S}_1(x)}\right)^{N_1(x,T-x)} \tag{8}$$

The parameter $p$, as well as the two survival functions $\widetilde{S}_i(x)$, *i=0,1* for genotypes, are unknown. Their values at each age x have to be estimated on the basis of the available data. Altogether, $2(X\text{-}x^*)+1$ parameters have to be estimated ($X\text{-}x^*$ parameters for each of two survival functions, plus $p$). Note that here we assume that $S_i(x^*)=1$, $i=0,1.$

**Demographic, Epidemiological and Other Constraints.** One cannot estimate the initial frequency $p$ and two survival functions $\widetilde{S}_i(x)$, *i=0,1* by maximizing (8) without additional conditions, because this model is non-identifiable. The demographic condition (5), where the function $\widetilde{S}(x)$, is known, allows us to reduce the number of model parameters to $X\text{-}x^*+1$, since (5) includes $X-x^*$ conditions (one for each age). Simulation studies show that these parameters are identifiable. Thus, the likelihood function (8) has to be maximized using constraint (8). Sometimes additional conditions can stem from the results of epidemiological studies in which the values of relative risks $r(x)$ for respective genotypes are estimated at some selected age interval. For example, the ratio of hazards for two genotypes,

$$\frac{\mu_0(x,T-x)}{\mu_1(x,T-x)} = r(x), \tag{9}$$

may be known for $x=x_1,x_2,\ldots x_n$ . In this case, the likelihood function (4.19) has to be maximized taking into account constraints (5) and (9). Hence, the number of model parameters becomes $X\text{-}x^*\text{-}n+1$. When the sample size of genetic data is large enough, the nonparametric (NP) method may provide acceptable estimates for survival functions $\widetilde{S}_i(x)$, *i=0,1* of genotypes and the initial frequency $p$. To take into account data from longitudinal surveys where multi-state model is used for description of changes in individual health status one will need to explore new model of population's heterogeneity.

**Applications.** In this section we will discuss the results of application of methods described above to the data.