Journal of Gerontology: BIOLOGICAL SCIENCES
2000, Vol. 55A, No. 7, B319–B328
Copyright 2000 by The Gerontological Society of America

Genes and Longevity: Lessons From Studies of Centenarians

A.I. Yashin,[1,2] G. De Benedictis,[3] J.W. Vaupel,[1,4] Q. Tan,[1] K.F. Andreev,[1] I.A. Iachine,[5] M. Bonafe,[6] S. Valensin,[6] M. De Luca,[3] L. Carotenuto,[8] and C. Franceschi[6,7]

[1]Max Planck Institute for Demographic Research, Rostock, Germany.
[2]Center for Demographic Studies and [4]Sanford Institute, Duke University, Durham, North Carolina.
[3]Departments of Cell Biology and [8]System Science, University of Calabria, Italy.
[5]Institute of Statistics and Demography, University of Southern Denmark, Odense.
[6]Department of Experimental Pathology, University of Bologna, Italy.
[7]Istituto Nazionale Riposo e Cura Anziani, Ancona, Italy.

In population studies of aging, the data on genetic markers are often collected for individuals from different age groups. The idea of such studies is to identify "longevity" or "frailty" genes by comparing the frequencies of genotypes in the oldest and in the younger groups of individuals. In this paper we discuss a new approach to the analysis of such data. This approach, based on the maximum likelihood method, combines data on genetic markers with survival information obtained from standard demographic life tables. This method allows us to evaluate survival characteristics for individuals carrying respective candidate genes. It can also be used in the estimation of the effects of allele–area and allele–allele interaction, either in the presence or absence of hidden heterogeneity. We apply this method to the analysis of Italian data on genetic markers for five autosomal loci and mitochondrial genomes. Then we discuss basic assumptions used in this analysis and directions of further research.

IN genetic studies of human aging and survival, the gene frequency (GF) method is often used (1,2). In this method, the contribution of candidate genes in the survival process is analyzed by comparing gene frequencies in two different age groups of individuals. According to strategies used currently to identify genes in multifactorial traits (3), allele pools are compared between sample groups of extremely old individuals (cases) and younger people (controls) from the same population. The observed case–control differences in allele frequencies are associated with the influence of a respective candidate gene on survival (4–12). To make proper classification on alleles as "frail," "robust," and "neutral," standard statistical methods, which identify differences in observed frequencies among case and control groups for different candidate alleles, are used. These methods, however, do not allow us to evaluate survival characteristics of individuals carrying candidate alleles. Because several alleles for a candidate locus are usually involved in an analysis, multiple comparisons based on the Bonferroni inequality, Scheffé's method, and others have often been used (13). The expediency of this procedure was, however, questioned by Rothman (14). Indeed, when carefully chosen hypotheses about specific alleles are tested, it does not seem reasonable to insist that each be adjusted for the mere presence of the other. However, when the presence of the effect in a selected locus is checked by a large number of tests, then correction for multiple comparison might be relevant. We (15) suggested the relative risk (RR) method of combining data on genetic markers with demographic information to obtain a more detailed characterization of genetic influence on survival and longevity. In this paper we extended the RR method to calculate the effects of hidden heterogeneity and interaction. We show that taking these effects into account can make the results of hypotheses testing significant after correction for multiple comparison. We discuss basic assumptions of this method and directions of further research.

Materials and Methods

Data

Five autosomal loci (APOB, REN, SOD1, SOD2, THO) and the mitochondrial locus (mtDNA) were considered. All these loci carry out biological functions that are expected to be crucial in successful aging and longevity. The APOB gene codes for apolipoprotein B, an exclusive protein of low-density lipoprotein (LDL), the main carrier of cholesterol in the blood. The REN locus codes for renin, an aspartylprotease that catalyzes the first step of the biosynthetic cascade leading to angiotensin 2. Both SOD1 and SOD2 code for superoxidodismutases that are involved in the elimination of superoxide radicals. The THO gene codes for tyrosine hydroxylase, the rate-limiting enzyme for the synthesis of catecholamines. Lastly, the mitochondrial genome contains genes for oxidative phosphorylation. The polymorphic systems were as follows: 3′APOB-VNTR [15 alleles (16)], HUMREN.4 [five alleles, (17)], SOD1-D21S223 [nine alleles (18)], $SOD2_{C/T}$ [C/T alleles (19)], HUMTHO.1 [six alleles (17, 20)], mtDNA haplogroups [nine alleles, (21)].

The data on genetic markers for the group of centenarians (aged 100 years and above) and the group of younger individuals (aged between 5 and 80 years) were obtained from samples collected in both Northern and Southern Italy. Altogether, 662 individuals were involved in the study; among them were 54 male and 143 female centenarians, and 220

Table 1. Italian Data on Genetic Markers by Sex and Area*

	Men	Women	Total
Younger Group			
South	145	158	303
North	75	87	162
Total	220	245	465
Centenarian Group			
South	28	60	88
North	26	83	109
Total	54	143	197
			662

*Genetic information comes from five autosomal genes, APOB, REN, SOD1, SOD2, THO, and mitochondrial DNA haplogroups.

male and 245 female younger individuals. Twenty-six male centenarians were from Northern Italy, and 28 were from Southern Italy. The number of female centenarians from the North was 83 and from the South 60. The younger group contained 75 men and 87 women from the North and 145 men and 158 women from the South. The distribution of this data by sex and area is shown in Table 1.

The ages of the subjects ranged from 5 to 109 years (the 5- to 19-year-olds were schoolchildren; the 20- to 29-year-olds were University undergraduate and graduate students; the subjects over 100 years old were gathered into a larger research project in progress in Italy; the others were volunteer donors). The samples used in this study were collected by eight institutions in Italy from 1995 to 1997. The ages of individuals in the centenarian group were verified by using information from demographic censuses, church registers, social security documents, and testimonies of relatives. The data for this group were aggregated with respect to age. For technical reasons, the number of individuals participating in the analysis of some loci is less than that mentioned above, so the number of observations for each gene varies (Table 2).

Relative Risk Method

The changes in gene frequencies with age within one cohort are produced by differences in hazard rates (risks of death) associated with the respective genes. This property suggests a new strategy for identifying frail and robust alleles. Instead of comparing gene frequencies between centenarians and younger individuals, one can evaluate and compare relative risks of death and survival distributions associated with different alleles by using the maximum likelihood method. These characteristics, however, cannot be identified without additional information on survival in respective age groups of individuals. Such information can be taken from standard demographic life tables. Observed risk factors such as geographic area of residence and sex may also be included in the likelihood function (15). The method of obtaining parameter estimates by maximizing the likelihood function of genetic data with demographic constraints, when mortality rates for individuals, carrying respective genotypes, are described by the Cox-type proportional hazard model (22), is called the relative risk method. Cox's model, widely used in the analysis of survival data, has proven to be a reliable tool for the evaluation of the influence of observed covariates on survival. The key assumption of this model is the multiplicative effects of influential factors on hazard rate. The theoretical aspects of this method in application to the analysis of survival data are investigated in an article by Cox (22). We suggested the use of Cox's model in the new approach to the analysis of cross-sectional data on genetic markers (15). This approach is a nontraditional one, because only censored information about life span of individuals with genetic markers is available. Note that because of the assumption about the proportionality of hazards, some details of genetic influence on survival may be lost if mortality rates for candidate genes or genotypes cross over. A comparison of methods used in genetic studies of centenarians has been done by us (23). A version of the RR method adjusted to our analysis is described in the Appendix.

In earlier research (15) we studied the effects of area, sex, and candidate allele on mortality and longevity by using the same sample of Italian data. The analysis of the allele–area and allele–allele interaction effects, performed in this paper, involves additional unknown parameters. To be able to produce reliable parameter estimates associated with interaction effects, we decided to aggregate data for men and women in this study and control only for regional differences. To show how data about area of residence and genetic markers may be included in the model, let us consider the hazard rate for an x-year-old individual for whom these data are available. Let us assume that the hazard rate for this individual may be represented in a Cox form as

$$\mu_0(x)e^{\sum_{i=1}^{3}\beta_i U_i}$$

(22). Here U_1 refers to the region (0 for the North and 1 for the South). Variable U_2 refers to the presence ($U_2 = 1$) or absence ($U_2 = 0$) of a candidate allele on a chromosome. Variable U_3 refers to the presence ($U_3 = 1$) or absence ($U_3 = 0$) of the same allele on the homologous chromosome, and $\mu_0(x)$ is an underlying hazard that characterizes the mortality rate for individuals with all $U_i = 0$. Thus, the survival function of an x-year-old individual can be represented as

$$S_0(x)^{r_1 r_2 r_3}.$$

Table 2. Sample Sizes Used for Each Polymorphic Locus*

Locus	Polymorphism	Alleles	Sample Size
APOB	3′ APOB-VNTR	31,33,35,37,39,41,43,45, 47,49,51,53,55	261
SOD1	D21S223	1,2,3,4,5,6,8,10	354
REN	HUMREN.4	7,8,10,11,12	295
THO	HUMTHO.1	6,7,8,9,10,11	520
SOD2	$(C/T)_{401nt}$	C,T	256
mtDNA	European haplogroups	H,I,J,K,T,U,V,W,X, others	372

*The allele nomenclature refers to the number of DNA repeats [APOB, REN, THO loci; see (11)]; to allele electrophoretic position (SOD1 locus); to C→T substitution [SOD2 locus; see (11)], and to restriction fragment length polymorphisms that define European mtDNA haplogroups [mtDNA; see (12)].

Here

$$r_i = e^{\beta_i U_i},$$

and

$$S_0(x) = e^{-\int_0^x \mu_0(u)du}.$$

In our study we assume an equal contribution of each allele in homologous chromosomes in the survival process. This assumption is natural when one cannot distinguish between the effects of homologous chromosomes. This yields $\beta_2 = \beta_3$, so the relative risk in individuals homozygous for the candidate allele is RR_2^2, where $RR_2 = e^{\beta 2}$ is the risk in individuals heterozygous for the candidate allele. This assumption allows us to use one unknown parameter (instead of two) to characterize the effect of a candidate allele on survival. The same effect may be achieved by the introduction of a covariate U which takes values 0, 1, and 2 (0 for the absence of a candidate allele in the locus, 1 for the presence of one candidate allele, and 2 for the presence of two candidate alleles). In case of such a description, the relative risk for individuals with the double copy of a given allele will be squared automatically. We prefer to use our description because it is more convenient for the representation of allele–allele interaction effects.

In the case of diploid (autosomal) loci, we have $N = 6$ groups. Each group is characterized by the area (0 for North and 1 for South), and one of three genotypes (0 for the absence of the candidate allele, 1 for the presence of this allele in one of two chromosomes, and 2 for the presence of this allele at both chromosomes). For the mtDNA locus we have $N = 4$ groups. It is convenient to represent the initial proportions of individuals in each of six groups in terms of two parameters: p_{0n} (the initial proportion of individuals from Northern Italy) and p_{0g} (the initial frequency of the candidate allele in a population). For example, let us consider one of the six groups for the APOB locus (say, group k). Assume that this number refers to the group of individuals from Southern Italy who have one candidate allele, say, APOB31. This group is characterized by the survival function $S_k(x) = S_0(x)^{RR_1 RR_2}$. The respective initial frequency will be $p_k = 2(1 - p_{0n})p_{0g}(1 - p_{0g})$; here the term $(1 - p_{0n})$ is the initial proportion of individuals from Southern Italy, and $2p_{0g}(1 - p_{0g})$is the initial proportion of heterozygous genotype in the case of Hardy–Weinberg equilibrium with one APOB31 allele. The other, say, group j of individuals from the North of Italy who have one APOB31 allele has the survival function $S_j(x) = S_0(x)^{RR_3}$ and the initial frequency $p_j = 2p_{0n}p_{0g}(1 - p_{0g})$. Similar representations can be written for each of the other four groups represented in likelihood (A1) in the Appendix and for the case of mtDNA.

Such representations for the initial proportions of individuals in the groups are based on two assumptions. The first is that the events of having a certain genotype with respect to a candidate allele and being a resident of Northern or Southern Italy for an individual are independent. The second is that the population of individuals in the study is in Hardy–Weinberg equilibrium. The first assumption is natural for populations in which genes under study are equally represented in all regions. This is usually the case in countries with high internal mobility and a relatively small area of residence. The second assumption is traditional in genetic studies. It provides a simple relationship between gene frequencies and genotype frequencies (3). Otherwise, one has to estimate five unknown initial frequencies for these groups. In principle, these assumptions are statistically testable. Unfortunately, the sample size of our data is not large enough to perform such testing.

Interaction Effects

To test for the interaction effects, we have to introduce the allele–area and allele–allele interaction terms in the survival model. In this new model the survival function $S_i(x)$, $i = 1,2,...6$ is given in terms of the Cox proportional hazard model with conditional hazards

$$\mu(x,U_1,U_2,U_3,U_4,U_5) = \mu_0(x)e^{\sum_{j=1}^{5}\beta_j U_i},$$

where U_1,U_2,U_3 are defined above, $\beta_2 = \beta_3$, and interaction variables U_4 and U_5 are respective combinations of U_1 (area), U_2 (allele), and U_3 (allele at homologous chromosome); that is, U_4 (area–allele, $U_4 = 1$ when $U_1U_2 = 1$, or $U_1U_3 = 1$, otherwise $U_4 = 0$), and $U_5 = U_2U_3$ (allele–allele).

Hidden Heterogeneity

Unobserved heterogeneity, also called frailty, is a major concern in a survival analysis, where individual differences cannot be safely ignored. To take hidden heterogeneity in mortality into account, we use the gamma-frailty model with mean 1 and variance σ^2 (24). In accordance with this model,

$$\mu(x,U_1,U_2,U_3,U_4,U_5,Z) = Z\mu_0(x,U_1,U_2,U_3,U_4,U_5). \quad (1)$$

The functional form of conditional survival function, $S(x,U_1,...,U_5)$, is derived in the Appendix. So, in addition to regression coefficients, β_i, and initial frequencies, p_{0n},p_{0g}, one has to estimate σ^2. The respective estimation procedure is called the HRR method.

RESULTS

Separate and Joint Analyses

First we performed a separate analysis of data on different alleles without taking interaction effects into account. Table 3 shows the results of these calculations.

The first column in this table characterizes the allele. The second shows the values of relative risks. The third and the fourth show the standard error for the estimated value of relative risk and the p value for testing the null hypothesis that relative risk is equal to 1 against the alternative that it is not, respectively. The last two columns show the estimates of the initial frequencies and their standard errors, respectively.

One can see from Table 3 that the relative risks of the area of residence are all greater than 1. This observation allows us to

Table 3. Estimated Risks and Initial Allele Frequencies for the Italian Data*

Allele	R	Standard Error	p Value	Initial Frequency	Standard Error
		Estimates for Alleles			
APOB31	1.089	.043	.042	.140	.015
D21S1	0.861	.079	.076	.025	.008
REN8	0.934	.036	.065	.731	.021
REN11	1.096	.045	.032	.160	.018
THO9	1.062	.034	.067	.221	.016
THO10	0.939	.030	.044	.199	.015
mtDNAhapl-V	0.789	.109	.052	.017	.008
mtDNAhapl-J	0.791	.057	.100	.028	.023
		Estimates for Area			
APOB31	1.095	.040	.016	.555	.030
D21S1	1.020	.041	.586	.462	.036
REN8	1.090	.046	.050	.421	.034
REN11	1.093	.046	.042	.423	.034
THO9	1.192	.040	.000	.662	.025
THO10	1.188	.040	.000	.661	.025
mtDNAhapl-V	1.088	.040	.027	.542	.030
mtDNAhapl-J	1.087	.040	.029	.541	.030

*The upper part of this table presents the estimates of relative risks and initial frequencies for the alleles whose relative risks differ significantly from ($p < .1$). The lower part of the table shows respective estimates for the area of residence (Southern with respect to Northern Italy). The following abbreviations are used: APOB31 for APOB-VNTR.31; D21S1 for D21S223.1; REN8 for HUMREN.8; REN11 for HUMREN.11; THO9 for HUMTHO.9; and THO10 for HUMTHO.10.

perform the joint analysis of data for all alleles by assuming that the effects of the area of residence are the same for all candidate alleles. The results of this analysis are shown in Table 4.

The first column in this table characterizes the allele. The second shows the values of relative risks for alleles. The third and the fourth show the standard errors for estimated value of relative risk and the *p* value for testing the null hypothesis that relative risk is equal to 1 against the alternative that it is not, respectively. The fifth and the sixth columns show the values of initial allele frequencies and their standard errors. The last two columns show the estimates of the initial frequencies for the area of residence and their standard errors, respectively. The likelihood ratio test confirms the legitimacy of the joint analysis.

Table 4. Estimates of Relative Risks and Initial Frequencies Using the RR Method*

Allele	r_{allele}	Standard Error	p Value	fr_{allele}	Standard Error	fr_{area}	Standard Error
APOB31	1.088	.043	.043	.139	.015	.561	.024
D21S1	0.865	.078	.086	.025	.008	.511	.027
REN8	0.932	.036	.057	.730	.021	.430	.027
REN11	1.096	.045	.032	.160	.018	.430	.027
THO9	1.064	.033	.057	.222	.016	.631	.021
THO10	0.939	.030	.040	.198	.015	.631	.021
mtDNAhapl-V	0.787	.109	.050	.018	.008	.552	.024
mtDNAhapl-J	0.844	.092	.090	.029	.010	.552	.024

*The model assumes the same risk for the area of residence for all alleles. Its estimate is 1.110 with a standard error of .015. The estimates of relative risks and initial gene frequencies for the alleles whose relative risks differ from 1 ($p < .1$) only are shown.

Table 5. Estimates of the Relative Risks, Initial Frequencies, and Interaction Terms for Allele–Area and Allele–Allele Interaction Using the RR Method*

Gene	Allele Frequency	Allele R	Allele p	Allele–Area r	Allele–Area p	Allele–Allele r	Allele–Allele p
APOB31	.140	1.124	.022	0.927	.204	—	—
D21S1	.024	0.824	.042	1.129	.318	—	—
REN8	.731	0.927	.047	0.958	.406	—	—
REN11	.161	1.122	.041	1.053	.521	0.807	.061
THO9	.221	0.985	.728	1.067	.230	1.477	.033
THO10	.196	0.900	.002	1.155	.004	—	—
mtDNAhapl-V	.019	0.813	.155	0.947	.722		
mtDNAhapl-J	.028	0.745	.011	1.367	.094		

*The risk of an area is taken to be the same for all alleles. Its estimate is 1.10 with a standard error of .015. In the last column, only cases with significant values of respective risks are reported. Shading indicates that allele–allele interaction is not possible for mitochondrial haplotypes.

Estimation of Interaction Effects

The results of a joint analysis of a model with interaction effects are shown in Table 5. One can see from Table 5 that the allele–area interaction is significant for the THO10 allele (*p* value = .004). This interaction increases the hazard rate of the THO10 carriers if they are residents of the South of Italy. Without such an interaction, THO10 is a robust allele. The THO9 allele becomes neutral in Table 5 with the *p* value changed from .057 in Table 4 to .728 in Table 5, but a significant allele–allele interaction effect is detected. This effect substantially increases the hazard of death for individuals with two THO9 alleles (homozygotes). In contrast, the allele–allele interaction for the REN11 allele has a robust effect. It reduces a carrier's death rate by a factor of 0.8, although the allele itself is classified as a frail allele.

The Effects of Hidden Heterogeneity

Note that as a way to minimize the number of unknown parameters, the same heterogeneity distribution for all eight alleles was assumed, and its estimation was done in a joint analysis of data. The highest value of likelihood is reached at $\sigma^2 = 0.66$ (see Figure 1).

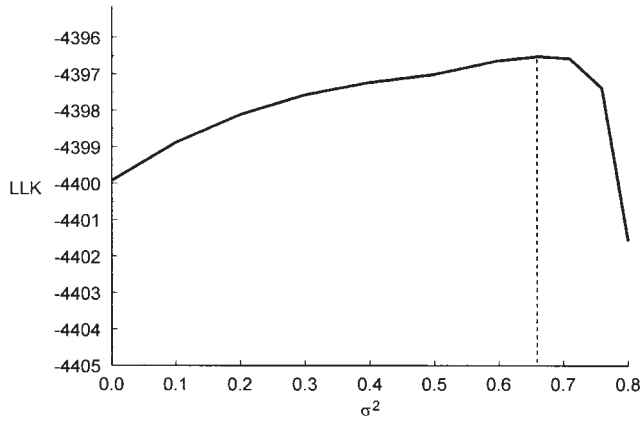

Figure 1. Graph of the profile of the log-likelihood (LLK) as a function of σ^2.

Table 6. Results of Estimation and Interaction Effects with Hidden Heterogeneity*

Gene	Allele Frequency	Allele		Allele–Area		Allele–Allele	
		R	*p*	*r*	*p*	*r*	*p*
APOB31	.140	1.480	.074	0.783	.222	—	—
D21S1	.023	0.504	.002	1.455	.401	—	—
REN8	.730	0.759	.022	0.862	.432	—	—
REN11	.163	1.483	.089	1.369	.393	0.481	.017
THO10	.192	0.686	.000	1.615	.017	—	—
mtDNAhapl-V	.017	0.492	.031	0.799	.589		
mtDNAhapl-J	.029	0.430	.000	2.716	.258		

*The gamma-distribution with mean 1 and variance σ^2 was used for heterogeneity variable. The estimate of $\sigma^2 = 0.66$. The risk of area is assumed the same for all alleles. Its estimate is 1.41 with a standard error of .014. Shading indicates that allele–allele interaction is not possible for mitochondrial haplotypes.

In calculation of the graph in Figure 1, the variance value changed manually, and the RR method was applied with each value of variance to estimate other parameters. Then the value of the likelihood function was calculated at each point. Because of the complicated structure of the likelihood function, this procedure was easier to perform than the direct likelihood maximization with respect to all parameters. With the larger data set, the heterogeneity parameters characterizing the frailty distribution for each candidate allele can, in principle, be estimated. The parameter estimates are shown in Table 6.

One can see from this table that taking heterogeneity into account makes estimates of all relative risks more distinct from 1. The estimates of interaction effects also changed. The *p* value for the allele–area interaction term for the TH10 allele increased from .004 (Table 5) to .017 (Table 6); however, the interaction effect is still significant. The estimates of all risks for the THO9 allele became nonsignificant, and they are not shown in the table. The significance of the allele–allele interaction term for the REN11 allele increased (the respective *p* value reduced from .061 to .017).

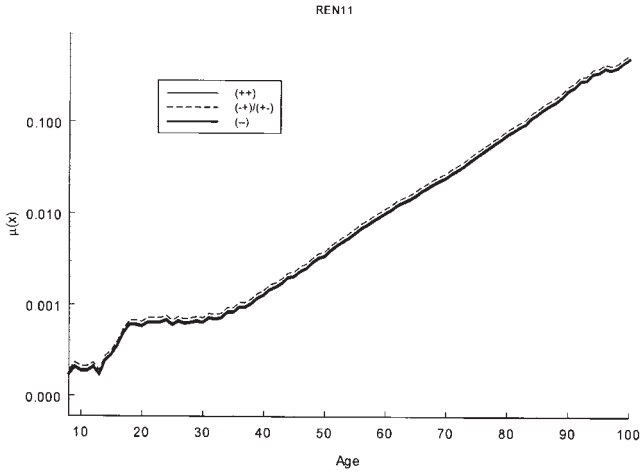

Figure 3. Age patterns of the estimated hazards for individuals carrying zero (thick solid line), one (dashed line), and two (thin solid line) copies of the REN11 allele in Italy. These estimates are obtained in the joint analysis of eight alleles without taking unobserved heterogeneity into account. The graph is shown in a logarithmic scale.

Figure 2 shows the age patterns of the estimated hazards for Italian individuals carrying zero, one, and two D21S1 alleles, respectively. These estimates are obtained without taking unobserved heterogeneity into account. Hazards are shown in a logarithmic scale. One can see from this figure that the hazard rate for carriers of the D21S1 allele is lower than that for noncarriers, so the D21S1 is a robust allele. All hazards have the same shape, as expected in the case of proportional hazard assumption. The estimates of hazard rates obtained in the case of separate and joint analyses practically coincide. Figure 3 shows the respective graph for the REN11 allele.

One can see from this graph that the REN11 is a frail allele. The graphs in Figures 4 and 5 show the estimates of the

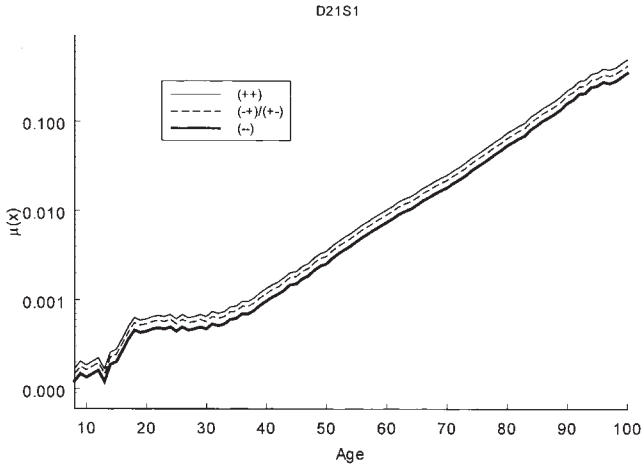

Figure 2. Age patterns of the estimated hazards for individuals carrying zero (thick solid line), one (dashed line), and two (thin solid line) copies of the D21S1 allele in Italy. These estimates are obtained in the joint analysis of eight alleles without taking unobserved heterogeneity into account. The graph is shown in a logarithmic scale.

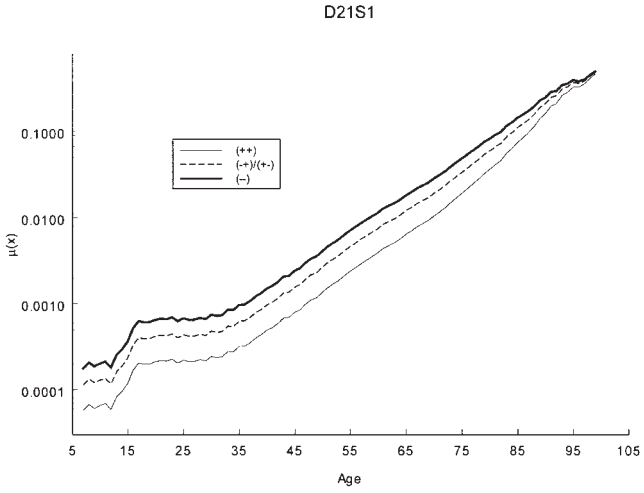

Figure 4. Age patterns of the estimated hazards for individuals carrying zero (thick solid line), one (dashed line), and two (thin solid line) copies of the D21S1 allele in Italy. These estimates are calculated by using a survival model with heterogeneity. The graph is shown in a logarithmic scale.

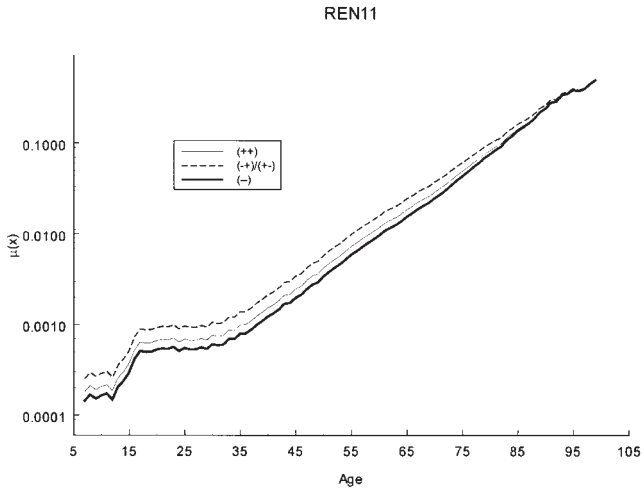

Figure 5. Age patterns of the estimated hazards for individuals carrying zero (thick solid line), one (dashed line), and two (thin solid line) copies of the REN11 allele in Italy. These estimates are calculated by using a survival model with heterogeneity. The graph is shown in a logarithmic scale.

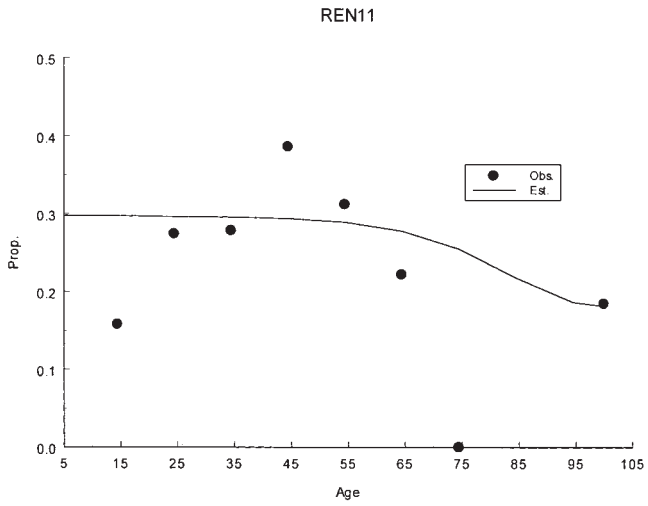

Figure 7. Graph of empirical (filled circles) and estimated (solid line) proportions of the carriers of at least one REN11 allele in Italy. The estimates correspond to the survival model with heterogeneity.

hazard rates for the same genotypes as in Figures 3 and 4, calculated with the heterogeneity model. One can see that the classification of D21S1 and REN11 remains the same. However respective mortality rates converge at old ages, as the frailty model predicts. Figures 6 and 7 show the graphs of empirical and estimated proportions of the carriers of D21S1 and REN11 alleles in Italy.

One can see that the proportion of the D21S1 allele increases with age, which indicates that it is a robust, or longevity, allele. The proportion of the REN11 allele declines with age, which indicates that it is a frailty allele. All calculations were done by using the GAUSS software package (25).

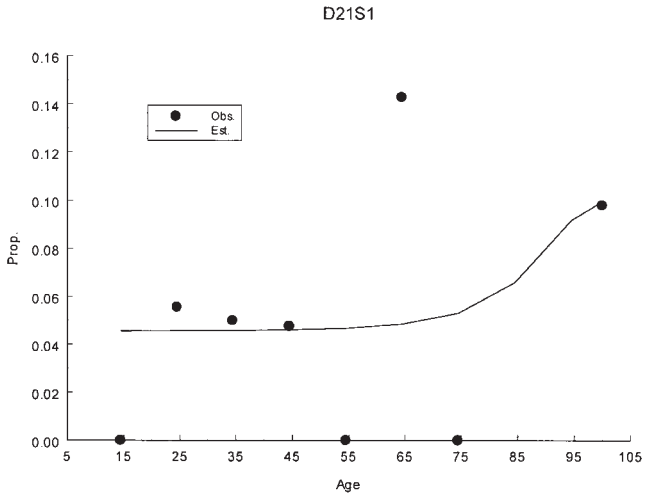

Figure 6. Graph of empirical (filled circles) and estimated (solid line) proportions of the carriers of at least one D21S1 allele in Italy. The estimates correspond to the survival model with heterogeneity.

DISCUSSION

Despite evident progress in establishing a connection between specific genes and longevity (26), some aspects of genetic studies for humans deserve additional efforts. In this paper we show that the use of demographic information together with data on genetic markers can substantially improve our knowledge about the role of genes in human mortality and longevity. The possibility of estimating hazard rates and survival functions for candidate genes opens a new avenue for the study of genetic effects on survival. Now, in addition to demographic information, the data and results of epidemiological studies can also be involved in the genetic analysis of longevity. This is because these studies often estimate values of relative risks or odds ratios for individuals of different genotypes at some age intervals. Such estimates must be consistent with the values of the hazard rates of genes or genotypes obtained in genetic studies of aging and longevity. The use of not only genetic but also epidemiological data, together with life-table demographic information, increases the power of the estimation procedures, expands the class of identifiable models, and permits us to address more sophisticated questions about roles of genes and environment in human mortality and longevity. Thus future studies will involve more additional information about direct and indirect influence of genes on health and survival.

When the sample size of genetic data is large enough, one can use nonparametric methods to estimate hazard rates or survival functions of genotypes from genetic and demographic data (23). In the case of a smaller sample size, the nonparametric estimates become unreliable, and semiparametric or parametric methods of data analysis can be used to improve the power of the estimation procedures. One has to realize, however, that the quality of approximation of real hazards may be compromised when parametric methods are used. The RR method discussed in this paper allows for the semiparametric estimation of the underlying hazards for re-

spective candidate genes or genotypes. This method allows us to control for the effects of observed covariates, interaction effects of genes, and gene–environment interactions. Because the assumption of proportionality of hazard at the entire age interval may be unrealistic, one can fit the RR model at smaller intervals and get a closer approximation of the real hazard rates for candidate alleles. In this case one has to be sure that the sample size of the data is large enough to get reliable parameter estimates at subintervals.

The RR method for combining genetic and demographic information without taking the effects of heterogeneity and interaction into account was suggested earlier (15). The idea to apply the frailty model to the analysis of data on genetic markers has also been discussed (23). This paper extends this approach to the analysis of allele–allele and allele–area interaction effects, as well as heterogeneity. The extended model contains more unknown parameters, and hence the power of the estimation procedure may be substantially reduced. To keep the number of unknown parameters at reasonable level, we did not distinguish between male and female mortality rates in this study. So in eight analyses shown in Table 3 and in the next Tables, the gender effect was not taken into account. However, we did control for regional differences. Tables 3 and 4 show the results of preliminary calculations without interaction effects. The comparison of results from Tables 3 and 4 with Table 1 in an eralier article (15) shows the difference in two detected alleles. The D21S6 allele is missed, and the allele THO9 is added in our study. This difference in the two studies may be attributed to the difference in the structure of respective models. All other alleles, qualified as frail and robust, are the same in both studies.

An analysis of the demographic situation in Italy shows that Italian survival presents some significant regional differences. In particular, the regions in Central Italy have a lower mortality than regions in Northern and Southern Italy. Northern Italy is less favorable for men. Southern Italy is less favorable for women (27). Calculations based on our data show a higher relative risk for the residents of Southern Italy. This may be the result of a higher proportion of women from Southern Italy in the control group and a lower proportion of such women in the centenarian group in our sample. Note also that for the effects of area and sex to be represented correctly, the data should be consistent with the demographic structure of respective populations. Otherwise, the estimates of these effects cannot characterize the population. Moreover, this inconsistency in the sample may bias the results of the genetic analysis. The use of Cox's regression model with explicitly represented covariates allows for the separation of the effects of genes from the effects of other influential factors on survival. However, the question of sensitivity of the estimates of genetic parameters to the changes in a composition of a sample of the data deserves special study.

The ability to take unobserved heterogeneity into account without changing the basic estimation procedure is an important advantage of the RR method. Such heterogeneity may exist as a result of the effects of other genes, or environmental factors not included in the analysis. The probability distribution of hidden heterogeneity is usually unknown, so its proper approximation is an important problem in a frailty modeling. The use of gamma-distributed frailty became popular because of its technical convenience, and because of its ability to explain deviant dynamics of mortality rate at old ages (24). Other distributions used in frailty modeling have also been discussed in the literature (28). A recent comparative analysis of several frailty distributions (29) used in genetic studies of susceptibility to death and longevity shows that gamma-frailty is a reasonable model for analyzing the effects of hidden heterogeneity in survival. Figure 1 shows the profile of the likelihood as a function of the variance of a frailty distribution calculated for a joint analysis of eight candidate alleles. A better strategy would be to estimate individual heterogeneity parameters for each candidate allele. Our attempts to estimate such parameters resulted in large standard errors. It is clear that more data are needed to realize this idea.

The analysis performed in this paper assumes that (i) the initial proportions of genotypes in all cohorts represented in a cross-sectional study are the same and that (ii) the survival functions of individuals carrying candidate alleles do not depend on the birth year of the cohorts. These assumptions were used in all earlier analyses of centenarian data cited in this paper. It is clear, however, that these assumptions are not realistic. In (23) we performed an analysis of sensitivity of the parameter estimates to the violations of these assumptions. The analysis shows that condition (i) is most sensitive to migration. This condition can be controlled by a historical demographic analysis of the data. The effects of differential total mortality between cohorts on observed gene frequencies [violation of assumption (ii)] depend on the patterns of allele–environment interaction. The details of such an interaction are unknown and cannot be estimated from the data used in our study. These effects may be small, if mortality rates for candidate alleles change proportionally, or large, if such interaction is more complex (23). Survival follow-up of individuals who provided genetic information will considerably strengthen the data.

The RR method developed in this paper is applied to several alleles in each of selected loci to test whether they are associated with longevity. For a given locus this procedure deals with a multiple testing, and care must be taken in order for the results not to be misinterpreted. Specifically, the need for adjustment for multiple comparisons depends on the question to be addressed by statistical analysis (30). Indeed, let us assume that one is interesting in testing the null hypothesis, H_0, that a selected allele, say APOB31, is a neutral one (i.e., whether the respective relative risk of death is equal to 1). Because only one test is applied to this allele, no adjustment for multiple comparison is needed.

If, however, one would like to test the null hypothesis, $H_{0\mathrm{APOB}}$, that all alleles in the APOB locus are neutral, then an adjustment for multiple comparison is needed. In this case one analyzes data on each of 15 different alleles available in this locus, and test 15 null hypotheses, H_{0j}, $j = 1,2,...15$, that the selected allele is a neutral one (i.e., whether the respective relative risk of death is equal to 1). Let α be a common significance level for each of these tests. It is clear that the significance level for testing the null hypothesis for the APOB locus, α_{APOB}, is related to α as follows:

$$\alpha_{APOB} = 1 - (1 - \alpha)^{15} \approx 15\alpha$$

One can see that in the case of high polymorphic loci, the value of α must be extremely small to provide an adequate significance level for testing the H_0 hypothesis that all alleles in the APOB locus are neutral. One can see from Tables 5 and 6 that taking into account interaction and heterogeneity effects decreases the p values for the estimates of risks associated with some alleles. For example, the p value for the THO10 allele was .04 in Table 4; then it became .002 in Table 5 when interaction effects were taken into account; and then it became less than .001 in Table 6 when, in addition, hidden heterogeneity was taken into account. This reduction in p values makes it possible to reject the null hypothesis for the THO locus (with six alleles) after correction for multiple testing.

The analysis of genetic data performed in this paper assumes that the mortality rate for some genes is lower (or higher) than that for the others at the entire demographic age interval. In this case the estimates of mortality rates for robust and frail alleles do not intersect, the estimates for proportions are monotone functions of age, and both GF and RR methods may be used for classification of alleles as robust, frail, and neutral. In addition, the RR method allows us to estimate the hazard rates and survival functions for respective candidate genes or genotypes. The empirical patterns of gene frequencies shown in Figures 4 and 5 suggest that the proportional hazard assumption may be too simplified and that the age trajectories of gene frequencies are not necessarily monotone functions of age. The reason for this may be an intersection of respective mortality curves for candidate genes. The presence of such intersections has been reported in studies (23,31) where other approaches to the analysis of genetic data on centenarians have been used. If hazard rates for genotypes intersect at a very young or at a very old age, the estimates of hazards calculated by the RR method still can approximate the average genetic effects on survival. However, in more complicated cases, important details related to the genetic regularities of aging process may be missed. For this reason the use of several approaches to the analysis of genetic data is recommended (23).

The intersection of hazard rates for carriers of different genes or genotypes suggests that survival to age 100 and more is not necessarily related to the presence of "robust genes," as it was generally believed before (5). Extended survival might be the result of a more sophisticated process of an organism's adaptation to the stresses of life. As part of this adaptation, genes responsible for a higher mortality at the beginning or in the middle of life may become beneficial at an advanced age (23). This effect may illustrate the important relationship among the ability to adapt, aging, and life span. Coping with the stresses of life, the organisms of individuals with disadvantageous genotypes are able to develop a higher "adaptation capacity" to the inevitable stresses of aging than those individuals with robust genotypes. If such adaptation mechanisms are in fact in effect, then the candidate genes also have to be searched for among those genes that produce a survival disadvantage earlier in life.

Acknowledgments

This research was partly (~ 40%) financed by the Italian Ministero Università Ricerca Scientifica Technologica (MURST) 1998–2000 project, "Longevity Determinants in Humans: the Model of Centenarians," by Instituto Nationale Riposo Cura Anziani (INRCA), Ancona (Italy), and by Grant PO1 AG08761-01 from the National Institutes of Health/National Institute on Aging. The authors are grateful to the reviewers, whose helpful comments improved the paper substantially. They also thank Cecilia Tomassini and Elisabetta Barbi for their qualified expertise concerning regional differences in Italian mortality, and Baerbel Splettstoesser and Karl Brehmer for help in preparing this paper for publication.

Address correspondence to Anatoli I. Yashin, Max Planck Institute for Demographic Research, Doberaner Strasse 114, 18057 Rostock, Germany. E-mail: yashin@demogr.mpg.de

References

1. Schächter F, Cohen D, Kirkwood T. Prospects for the genetics of human longevity. *Hum Genet.* 1993;91:519–526.
2. De Benedictis G. Genes and longevity. *Aging Clin Exp Res.* 1996;8: 367–369.
3. Falconer DS. *Introduction to Quantitative Genetics.* New York: Wiley; 1989.
4. Proust J, Moulias R, Fumeron F, et al. HLA and longevity. *Tissue Antigens.* 1982;19:168–173.
5. Takata H, Suzuki M, Ishii T, Sekiguchi S, Iri H. Influence of major histocompatibility complex region genes on human longevity among Okinawan-Japanese centenarians and nonagenarians. *Lancet.* 1987;ii: 824–826.
6. Kervinen K, Savolainen MJ, Sallokannen J, et al. Apoliporotein E and B polymorphisms—longevity factors assessed in nonagenarians. *Atherosclerosis.* 1994;105:89–95.
7. Louhija J, Miettinen HE, Kontula K, Tikkanen MJ, Miettinen TA, Tilvis RS. Ageing and genetic variation of plasma apolipoproteins. Relative loss of the apolipoprotein E4 phenotype in centenarians. *Artherioscler Thromb.* 1994;14:1084–1089.
8. Schächter F, Faure-Delanef L, Guenot F, et al. Genetic association with human longevity at the APOE and ACE loci. *Nature Genet.* 1994;6:29–32.
9. De Benedictis G, Falcone E, Rose G, et al. DNA multiallelic systems reveal gene/longevity associations not detected by diallelic systems. The APOB locus. *Hum Genet.* 1997;99:312–318.
10. De Benedictis G, Carotenuto L, Carrieri G, et al. Age-related changes of the 3′APOB–VNTR genotype pool in ageing cohorts. *Ann Hum Genet.* 1998;62:115–122.
11. De Benedictis G, Carotenuto L, Carrieri G, et al. Gene/longevity association studies at four autosomal loci (REN,THO, PARP, SOD2). *Europ J Hum Genet.* 1998;6:534-541.
12. De Benedictis G, Rose G, Carrieri G, et al. Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. *FASEB J.* 1999;13:1532–1536.
13. Hsu JS. *Multiple Comparisons.* New York: Chapman and Hall; 1996.
14. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology.* 1990;1:43–46.
15. Yashin AI, Vaupel JW, Andreev KF, et al. Combining genetic and demographic information in population studies of ageing and longevity. *J Epidemiol Biostat.* 1998;3:211–216.
16. Boerwinkle E, Xiong W, Fourest E, Chan L. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3′ hypervariable region. *Proc Natl Acad Sci USA.* 1989;86:212–216.
17. Edwards A, Hammond HA, Jin L, Caskey T, Chakraborty R. Genetic variation at five trimeric and tertameric tandem repeat loci in four human population groups. *Genomics.* 1992;12:241–253.
18. Rosen DR, Sapp PC, O'Regan J, et al. Dinucleotide repeat polymorphisms (D21S223 and D21S224). *Hum Mol Genet.* 1992;1:547.
19. Rosenblum JS, Gilula NB, Lerner RA. On signal sequence polymorphisms and diseases of distribution. *Proc Natl Acad Sci USA.* 1996;93: 4471–4473.
20. Puers C, Hammond HA, Jin L, Caskey T, Schumm JW. Identification of repeat sequence heterogeneity at the polymorphic short tamdem repeat locus HUMTHO1 (AATG)*n* and reassignment of alleles in population analysis by using a locus specific allelic ladder. *Am J Hum Genet.* 1993;53:953–958.
21. Torroni A, Huoponen K, Francalacci P, et al. Classification of European mtDNAs from an analysis of three European populations. *Genetics.* 1996;144:1835–1850.

22. Cox DR. Regression models and life-tables (with discussion). *J Roy Statist Soc B.* 1972;34:187–220.
23. Yashin AI, De Benedictis G, Vaupel J, et al. Genes demography and life span: the contribution of demographic data in genetic studies of aging and longevity. *Am J Hum Genet.* 1999;65:1178–1193.
24. Vaupel JW, Yashin AI. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *Am Statistician.* 1985;39:176–185.
25. *GAUSS: Mathematical and Statistical System.* Vol. I: System and Graphics Manual. Maple Valley, WA: Aptech Systems; 1996.
26. Jazwinski SM. Longevity, genes and aging. *Science.* 1996;273:54–59.
27. Population projections by sex, age and region. Base 1.1. (Previsioni della popolazione residente per sesso, eta e regione Base1.1.) Roma: Sistema Statistico Nationale, Instituto Nazionale di Statistica; 1996.
28. Aalen OO. Effects of frailty in survival analysis. *Statist Meth Med Res.* 1994;3:227–243.
29. Yashin AI, Begun AZ, Iachine IA. Genetic factors in susceptibility to death: comparative analysis of bivariate survival models. *J Epidemiol Biostat.* 1999;4:53–60.
30. Cox DR. A remark on multiple comparison methods. *Technometrics.* 1965;7:223–224.
31. Toupance B, Godelle B, Gouyon P-H, Schachter, F. A model for antagonistic pleiotropic gene action for mortality and advanced age. *Am J Hum Genet.* 1998; 62:1525–1534.

Received June 17, 1999
Accepted November 10, 1999
Decision Editor: Jay Roberts, PhD

Appendix

The Likelihood Function

Let $\pi_i(x)$, $i = 1,2,...N$ be the proportion of x-year-old individuals from the ith group in some cross-sectional study performed in year T, and let N_{ix} be the respective numbers of individuals observed in this study. Then the likelihood function of the data is

$$L \sim \prod_{x=x_0}^{X} \prod_{i=1}^{N} \pi_i(x)^{N_{ix}}, \tag{A1}$$

where

$$\pi_N(x) = 1 - \sum_{i=1}^{N-1} \pi_i(x),$$

and

$$\pi_i(x) = \frac{p_i S_i(x)}{\sum_{j=1}^{N} p_j S_j(x)}. \tag{A2}$$

Here p_i represents the initial proportion of individuals from group i. Note that if data start from age x_0, then p_i denotes respective proportion at this age. For each $i = 1,2,...6$, the survival function $S_i(x)$ is represented in terms of the Cox (22) proportional hazard model with conditional hazards

$$\mu(x,U_1,U_2,U_3) = \mu_0(x) e^{\sum_{j=1}^{3} \beta_j U_j}$$

(we assume here that $\beta_2 = \beta_3$) and with the respective combination of values for U_1, U_2, and U_3.

Estimation Procedure

The likelihood, Equation (1), must be maximized with respect to parameters p_{0n}, p_{0g} and risks $RR_i = e^{\beta_i}$, $i = 1,2$ under the constraint

$$S(x) = \sum_{j=1}^{N} p_j S_j(x). \tag{A3}$$

Here the values of survival functions $S(x)$ are taken from the official demographic life tables for the Italian population for 1996. The values of $S_j(x)$ depend on

$$S_0(x) = e^{-\int_0^x \mu_0(u)du},$$

and $RR_i = e^{\beta_i}$, $i = 1,2$. The estimation procedure, which takes into account constraint (A3), starts with the maximization of likelihood (A1) with respect to initial proportions p_{0n}, p_{0g} and risks RR_i, $i = 1,2,3$, taking the initial guess of $S_0(x)$ to be equal to, say, $S(x)$ (which is a known function of x). Then the estimates of p_{0n}, p_{0g} and risks RR_i, $i = 1,2,3$ are substituted into Equation (A3), from which the second guess of $S_0(x)$ is calculated. This guess is substituted in Equation (A1) with unknown parameters, p_{0n}, p_{0g}, and RR_i, $i = 1,2,3$. Then likelihood (A1) is maximized again to produce a second guess of these parameters, and the procedure is repeated until convergence occurs.

The Likelihood Function in the Case of Joint Analysis

The joint analysis of data for several candidate alleles makes sense when some parameters in eight likelihood functions are the same. In our analyses we assume the same risk for the area of residence R_1 (i.e., regression coefficient β_1 is the same for all candidate genes). In this case the likelihood function of joint data is

$$L \approx \prod_{k=1}^{n} \prod_{x=x_0}^{X} \prod_{i=1}^{N} \pi_i^k(x)^{N_{ix}^k}. \tag{A4}$$

Here

$$\pi_N^k(x) = 1 - \sum_{i=1}^{N-1} \pi_i^k(x),$$

$n = 8$, and

$$\pi_i^k(x) = \frac{p_i^k S_i^k(x)}{\sum_{j=1}^{N} p_j^k S_j^k(x)}. \tag{A5}$$

Here p_i^k represents the initial proportion of individuals in group i for the k_{th} candidate allele.

Survival in Heterogeneous Population

Let $\mathbf{U}$ denote vector $U_1,U_2,...,U_5$, and let $S(x|Z,\mathbf{U}) = S(x,U_1,U_2,...,U_5,Z)$ be conditional the survival function corresponding to hazard (1):

$$S(x|Z,\mathbf{U}) = \exp\{-ZH(x,\mathbf{U})\}, \tag{A6}$$

where

$$H(x,\mathbf{U}) = e^{\sum_{i=1}^{5} \beta_i U_i} \int_0^x \mu_0(s)ds.$$

Let frailty Z be gamma distributed with mean 1 and variance σ^2. Then

$$S(x|\mathbf{U}) = \int_0^\infty \frac{e^{-zH(x,\mathbf{U})} z^{k-1} \lambda^k e^{-\lambda z}\, dz}{\Gamma(k)}. \quad \text{(A7)}$$

here

$$\frac{(z^k \lambda^k e^{-\lambda z})}{[\Gamma(k)]}$$

is the probability density function for gamma distribution with the shape parameter k, and the scale parameter λ, and $\Gamma(k)$ denotes a gamma function. The mean of this distribution is

$$\frac{k}{\lambda},$$

and the variance

$$\sigma^2 = \frac{k}{\lambda^2}.$$

To calculate equation (A7), note that this equation can be rewritten in the form

$$S(x|\mathbf{U}) = [\lambda + H(x,\mathbf{U})]^{-k} \lambda^k \int_0^\infty \frac{[\lambda + H(x,\mathbf{U})]^k z^{k-1} e^{-[\lambda + H(x,\mathbf{U})]z}}{\Gamma(k)} dz, \quad \text{(A8)}$$

and because the integral in Equation (A8) is equal to 1 we get

$$S(x|\mathbf{U}) = \lambda^k [\lambda + H(x,\mathbf{U})]^{-k}.$$

Taking into account that

$$\frac{k}{\lambda} = 1,$$

and hence

$$\sigma^2 = \frac{1}{\lambda} = \frac{1}{k},$$

we get

$$S(x|\mathbf{U}) = \left[1 - \sigma^2 \ln S_0(x) e^{\sum_{i=1}^{5} \beta_i U_i}\right]^{-\frac{1}{\sigma^2}}. \quad \text{(A9)}$$

This survival function is used to characterize proportions of individuals in respective groups in the likelihood function.