Published in final edited form as:

Biogerontology. 2007 June ; 8(3): 243–255.

The accuracy of statistical estimates in genetic studies of aging can be significantly improved

Anatoli I. Yashin, Konstantin G. Arbeev, and Svetlana V. Ukraintseva

Center for Demographic Studies, Duke University, 2117 Campus Drive, Box 90408, Durham, NC 27708-0408, USA

Abstract

The sample size of the data used in genetic studies is often a factor limiting the accuracy of statistical estimates. In this paper we suggest a new approach to evaluation of genetic influence on risk of development of aging-related health disorders. The approach results in substantial improvement of the accuracy of statistical estimates without an increase in the size of the genetic sample. The approach is based on the joint analysis of data from the genetic samples and easily accessible non-genetic data, such as data collected in epidemiological, demographic, and longitudinal studies of human aging and aging-related pathologies.

Keywords

Accuracy of estimates; Aging; Demographic data; Gene; Longitudinal data; Model

Introduction

The role of genetic and non-genetic factors—as well as gene–environment interaction in the risks of aging-related diseases, disability, and longevity—would be much better understood if one could compare age patterns of respective hazard (incidence, disability, mortality) rates for carriers of selected alleles (genotypes) evaluated under various conditions (De Benedictis et al. 2001; Todd 2001). Such understanding is important for predicting the health and survival consequences of occupational exposure, environmental pollution, natural disasters, or technological accidents. It is also crucial for the development of drugs to protect the human organism from damage caused by unexpected exposures to harmful conditions. The traditionally cited reason for limited knowledge about the role of genetic and non-genetic factors in risks of developing age-related health disorders is the inadequate sample size of available genetic data, i.e. the number of individuals from whom genetic data were taken. Since an increase in the number of such individuals is an expensive business, the sample size is always considered to be a factor limiting the desirable accuracy of genetic estimates (Hoenig and Heisey 2001). That is why finding an indirect way of increasing this accuracy is an important and challenging task.

In this paper we suggest a new approach to genetic analysis and show that substantial improvement in the accuracy of statistical estimates can be reached without increasing the sample size of most expensive genetic data via a statistical estimation procedure that involves a joint analysis of genetic and appropriate non-genetic data. Fortunately, the necessary non-genetic data are easily accessible. These include data from epidemiological, demographic, and longitudinal studies of human aging and age-associated disorders. We have applied the new method to the joint analysis of genetic data from a sub-sample of the 1999 National Long Term

Care Survey (NLTCS) and of other (survey) data collected in the NLTCS. The data for genetic analysis were obtained from Dr. George Martin's lab, the University of Washington School of Medicine. To illustrate the advantage of combining such data, we performed an extensive simulation study using values of incidence rates for carriers and non-carriers of a hypothetical allele comparable to those calculated from the real data.

Methods and data

Which data have to be combined?

It seems intuitively clear that when the size of the genetic sample is fixed, the accuracy of statistical estimates can be improved only if some additional information (data) about the influence of genetic or non-genetic factors on morbid risks is taken into account in an appropriate way. What kind of data could they be? It turns out that data containing useful genetic information are not only available, but can be easily accessed. Here we consider the effects of joint analysis of three types of data. One is the main genetic sample. The most simple and straightforward way of evaluating the genetic influence on risks of developing aging-related disorders is the direct calculation of respective hazard rates from the data by following up carriers of selected alleles (genotypes) who participated in a cross-sectional sample. The empirical estimates of risks are then compared and, if accuracy permits, a conclusion about the strength of genetic effects on the risk of developing a selected disorder is drawn. In many practical cases, however, such a conclusion cannot be made, because empirical estimates of the risk of getting the disease (the hazard rate) are unreliable. An increase in the sample size of such genetic data is the straightforward way of coping with the problem of lack of accuracy. This strategy, however, requires recruitment of new individuals and collection of blood samples or other biological specimens, as well as performing a genetic analysis. Its restricted application is related to the substantial costs involved in all these procedures.

The genetic data of the second type are given in age trajectories of genetic frequencies. Such trajectories, often observed in cross-sectional genetic studies, are an important source of information on genetic influence on the risk of disease development. Indeed, the difference in genetic frequencies among groups of individuals of different ages could result from differential selection. The proportion of healthy carriers of respective alleles (genotypes) having higher hazard rates declines with age, and the proportion of those with lower hazard rates increases. Note that data on proportions (prevalence) of selected alleles among healthy individuals or among survivors in different age groups are available much more often than data on respective hazards. For example, these data are typical for genetic association studies (Varcasia et al. 2001; Carrieri et al. 2001; Cardon and Bell 2001). This is because prevalence data are the immediate results of a cross-sectional study. To obtain data on incidence, one has to follow up the respective population for at least one more year. Such a follow up requires additional investment, and that investment often is not possible. The methodological challenge in this situation is to develop appropriate statistical methods and approaches capable of releasing hidden information on the genetic influence on risks of selected diseases from data on age patterns of prevalence. Ideas and approaches to combining data on genetic frequencies with demographic information are already discussed in the literature (Yashin et al. 1999,2000; Lewis and Brunner 2004).

The third type of data characterizes marginal hazards. Such data or the estimates of the age trajectories of marginal risks (without genetic specification) for selected diseases are often available from epidemiological, demographic or other studies. The values of marginal hazards contain hidden information about the genetic composition of a respective population that consists of several groups of individuals—carriers of different alleles (genotypes). Hence this population is a discrete mixture of genetically different groups of individuals who are characterized by initial proportions and age patterns of incidence and mortality rates

characterizing risks of disease development or death in respective groups. The important feature, which makes our analysis possible, is that risks of disease development for individuals in the sample selected for the genetic analysis and in such a heterogeneous population experience influence of the same genetic and non-genetic factors. Therefore, the overall age patterns of hazard rates describing the dynamics of health deterioration in such a population contain important information about the genetic influence on chances of getting a disease or becoming disabled, although specific information about such influence is hidden (not observed directly).

Likelihood function for joint analysis of genetic and non-genetic data

The maximum likelihood method can be used to estimate parameters in the model for the joint analysis of genetic and non-genetic data. The likelihood function consists of the following three components associated with specific data sets described above.

Likelihood function for genotype-specific health transitions—The population under study consists of carriers ('1') and non-carriers ('0') of selected allele (genotype). A cross-sectional study performed in the year T , deals with $N(x, T-x)$ individuals in the age group x , $x = x^*, x+1, x+2, \dots, X$. Let $N_0(x, T-x)$ be the number of non-carriers and $N_1(x, T-x) = N(x, T-x) - N_0(x, T-x)$ the number of carriers of respective allele (genotype) detected in the genetic study. Denote by $M_0(x, T-x)$ and $M_1(x, T-x)$ the numbers of non-carriers and carriers of this allele (genotype) in the age group x , who contracted a selected disease between years T and $T+1$. The estimates of incidence rates $\mu_i(x, T-x)$, $i = 0, 1$, for non-carriers and carriers of respective allele (genotype) can be obtained from maximization of binomial likelihood function:

$$L_1 \sim \prod_{x=x^*}^X \prod_{i=0}^1 \mu_i^{M_i(x, T-x)} (1 - \mu_i)^{N_i(x, T-x) - M_i(x, T-x)} \quad (1)$$

In currently available samples of the data, such estimates have large standard errors in each age group. When age trajectories of $\mu_i(x, T-x)$ are described by some parametric functions with unknown parameters, the likelihood (1) can be used for estimation of these parameters.

Likelihood function for frequencies of genotypes—Let $\pi(x, T-x)$ be the frequency of non-carriers of some allele (genotype) at age x in the cohort born in the year $T-x$. The numbers $N_0(x, T-x)$ and $N_1(x, T-x)$, $x = x^*, x^*+1, \dots, X$, can be considered as a result of $N(x, T-x) = N_0(x, T-x) + N_1(x, T-x)$ trials from the binomial distribution with the probability of success $\pi(x, T-x)$. The estimates of these frequencies can be obtained from maximization of the binomial likelihood function:

$$L_2 \sim \prod_{x=x^*}^X \pi^{N_0(x, T-x)} (1 - \pi)^{N_1(x, T-x)} \quad (2)$$

When age trajectories of $\pi(x, T-x)$ are described parametrically, the likelihood (2) can be used for estimation of these parameters. This property is used in the estimation procedure that combines different data sets.

The joint analysis of data on genotype-specific health transitions collected in years of follow up of individuals from the genetic sub-sample and cross-sectional data on the number of individuals carrying the selected genotype can be performed using a joint likelihood function

$L = L_1 L_2$ taking into account the connection between $\pi(x, T-x)$, $\mu_i(x, T-x)$ and the aggregate incidence rate $\bar{\mu}(x, T-x)$ described below.

Likelihood function for the hazard rate in a mixture of carriers and non-carriers

—A substantial amount of genetic information is hidden in aggregated data on health transitions at different ages collected in longitudinal studies. These data are informative because they represent health transitions in a population, which is a mixture of carriers and non-carriers of respective alleles (genotypes). For example, the aggregated incidence rate $\bar{\mu}(x, T-x)$ measured in longitudinal studies is a mixture of two incidence rates for carriers and non-carriers of respective genotypes:

$$\mu(x, T-x) = \pi(x, T-x)\mu_0(x, T-x) + (1 - \pi(x, T-x))\mu_1(x, T-x). \quad (3)$$

This additional relationship is taken into account in the joint analysis of data on frequencies of genotypes, genotype-specific hazards and aggregated hazards. For some diseases, the estimate of $\bar{\mu}(x, T-x)$ can be known from other studies. In this case Eq. (3) can be used as a constraint in maximization of the likelihood function $L = L_1 L_2$.

In case when $\bar{\mu}(x, T-x)$ can be calculated from data on longitudinal studies (such as in our applications to the genetic and non-genetic data from the 1999 NLTCs), one can perform a joint analysis of data using the likelihood function $L = L_1 L_2 L_3$, where L_3 is the likelihood corresponding to the longitudinal data (i.e., aggregated data on the incidence in a mixture of carriers and non-carriers):

$$L_3 \sim \prod_{x=x^*}^X \mu(x, T-x)^{N^*(x, T-x)} (1 - \mu(x, T-x))^{N(x, T-x) - N^*(x, T-x)}. \quad (4)$$

Here $N^*(x, T-x)$ is the number of individuals contracted the selected disease during the interval $(T, T+1)$, and $N(x, T-x)$ is the number of individuals at risk at the beginning of this interval.

The frequencies $\pi(x, T-x)$ can be expressed in terms of survival functions for carriers and non-carriers, $S_i(x, T-x) = \exp(-\int_{x^*}^x \mu_i(t, T-t)dt)$ $i = 0, 1$, and the aggregate survival function (survival in a mixture of carriers and non-carriers), $S(x, T-x) = pS_0(x, T-x) + (1-p)S_1(x, T-x)$ where $p = \pi(x^*, T-x^*)$ as follows:

$$\pi(x, T-x) = pS_0(x, T-x) / S(x, T-x). \quad (5)$$

Expressions (3) and (5) indicate the key property of the method that makes the increase in the accuracy of estimates in the joint analysis of genetic and non-genetic samples possible: the components of the likelihood function related to genotype-specific hazard rates (L_1 , see (1)), frequencies of genotypes (L_2 , see (2)), and the aggregated hazard rate in a mixture of carriers and non-carriers in the ‘non-genetic’ sample (L_3 , see (4)) depend on the same parameters (those of $\mu_i(x, T-x)$ and the initial proportion p). Therefore, the estimates of respective parameters in the joint analysis are based on a substantially larger sample (‘non-genetic’) compared to the traditional analysis based on a smaller genetic sample.

If difference in hazard rates between subsequent cohorts is small, then the likelihood L_3 in (4) can include additional follow up data on health transitions at intervals $(T+1, T+2)$, $(T+2, T+3)$, etc. In the presence of time trends in hazard rates (or initial proportions), one can use parametric models describing the time dependence of hazards (proportions) and estimate respective parameters from the joint likelihood.

Simulation study

We performed a simulation study to illustrate to what extent the joint analysis of genetic and non-genetic data improves the accuracy of parameter estimates compared to the analysis of genetic data set alone. In this study we used the results of our preliminary analysis of NLTCS data to make simulated, realistic-looking hazard rates. In the simulation procedure, we modeled a mixture of carriers and non-carriers of a hypothetical allele/genotype with hazard rates (i.e., incidence of some disease or death) among non-carriers ($i = 0$) and carriers ($i = 1$) described by the Gompertz's functions $\mu_i(x) = a_i e^{b_i(x-65)}$ with the following parameter values: $a_0 = 0.008$, $b_0 = 0.08$, $a_1 = 0.005$, $b_1 = 0.13$. The initial proportion of non-carriers was $p = 0.75$. For each individual, we simulated the age at death (or contracting a disease) according to the respective Gompertz hazard rate. We simulated such data for 20,000 individuals (5,000 carriers, 15,000 non-carriers). This group represents the 'non-genetic' part of data, i.e., data on the allele/genotype are not available for these individuals. This group was used to calculate the number of individuals at risk and the number of deaths (or contracting a disease) at different ages in a mixture of carriers and non-carriers. Then we randomly selected 500 out of 20,000 individuals as a 'genetic' sample. For this group, data on the allele/genotype are available. This group was used to calculate the number of individuals at risk and the number of deaths (or contracting a disease) at different ages for carriers and non-carriers of the allele/genotype. Such sample sizes are comparable to the actual numbers of individuals in the 1999 NLTCS from whom the blood was collected, and the number of individuals who participated in the screening survey of each wave of the NLTCS. Such a procedure was repeated 1,000 times to produce 1,000 simulated data sets.

We estimated parameters of Gompertz hazard rates separately in each data set using two models. The first one used a traditional approach to the analysis of genetic data, which is based on the genetic sample alone (500 individuals). The respective hazard rates for carriers and non-carriers were estimated using the likelihood function

$$L_1 = \prod_{x=x^*}^X \frac{1}{\prod_{i=0}^1 \mu_i(x)^{M_i(x)} (1 - \mu_i(x))^{N_i(x) - M_i(x)}}, \quad (6)$$

where $N_i(x)$, $i = 0, 1$, are the numbers of non-carriers and carriers survived to age x (or did not contract a disease to this age), and $M_i(x)$, $i = 0, 1$, are the numbers of non-carriers and carriers died (or contracted a disease) at age x . The parameters estimated in this model are a_0 , b_0 , a_1 , and b_1 . The above likelihood function was maximized separately for each data set to produce 1,000 estimates of parameters a_0 , b_0 , a_1 , and b_1 . Respective results are shown in Table 1 and Fig. 1. To evaluate the sample size of genetic data needed to obtain the accuracy of estimates comparable to that reached in the joint analysis of genetic and non-genetic data, we also estimated the entire sample of 20,000 individuals using the likelihood (6). Respective results are shown in Table 1 and Fig. 2.

The second model uses the joint analysis of 'genetic' (500 individuals) and 'non-genetic' (20,000 individuals) samples. The respective hazard rates for carriers and non-carriers were estimated using the joint likelihood function $L = L_1 L_2 L_3$, where L_1 is the likelihood for the 'genetic' sample (6), L_2 is the likelihood for frequencies $\pi(x)$:

$$L_2 = \prod_{x=x^*}^X \pi(x)^{N_0(x)} (1 - \pi(x))^{N_1(x)}, \quad (7)$$

and L_3 is the likelihood for the 'non-genetic' sample:

$$L_3 = \prod_{x=x^*}^X \mu(x)^{N^*(x)} (1 - \mu(x))^{N(x) - N^*(x)}, \quad (8)$$

where $N^*(x)$ is the number of individuals in the non-genetic sample who died (contracted a disease) at age x , $N(x)$ is the number of individuals at risk at this age, and $\tilde{\mu}(x)$ is the hazard rate in the non-genetic sample (a mixture of carriers and non-carriers of the allele/genotype): $\tilde{\mu}(x) = \pi(x)\mu_0(x) + (1 - \pi(x))\mu_1(x)$. The parameters estimated in this model are a_0, b_0, a_1, b_1 and p . The likelihood function L was maximized separately for each data set to produce 1,000 estimates of respective parameters. Respective results are shown in Table 1 and Fig. 1.

We also simulated data for the cross-sectional sample with a time trend in the initial proportions of allele/genotype. We assumed a linear trend in the initial proportions over time, $p(x) = p_0 + q_0(x - x^*)$, where x denotes cohorts and x^* is the youngest cohort, $p_0 = 0.75$, and $q_0 = -0.0025$. We also assumed that hazard rates $\mu_i(x)$ are equal in different cohorts. We run the simulation algorithm described above for each cohort with numbers of carriers and non-carriers corresponding to the cohort-specific values $p(x)$ and collected 1,000 data sets of 20,000 individuals in non-genetic samples and 500 individuals in genetic samples. We estimated each data set using three models. The first model is based on the genetic sample and the likelihood (1). The second model uses the joint analysis with the likelihood (1), (2), (4) with fixed initial proportions p . The third model uses the joint analysis with the likelihood (1), (2), (4) with a linear trend in initial proportions. Respective results are shown in Table 3.

Application to the NLTCS data

We applied this approach to the analysis of data from a genetic sub-sample of the 1999 NLTCS data. The 1999 NLTCS contains information on 19,907 individuals aged 65+. Data on *APOE* alleles were collected for 1,805 participants of the 1999 NLTCS for both sexes. Altogether, 420 carriers and 1,385 non-carriers of *APOE* e4 allele were observed. All NLTCS records were linked to Medicare claims data for 1982–2001. The following diseases and ICD-9-CM codes were used to define individuals with cognitive disorders: Alzheimer's disease (AD)—331.0; mental disorders due to organic brain damage (no AD)—290.4, 310, 438.0; Parkinson's disease—332; presenile dementia—290.1; senile dementia—290.0, 290.2, 290.3, 290.8, 290.9, 331.2. Using the Medicare records, we evaluated events related to cognitive disorders occurring in years 2000 and 2001. Note that merging NLTCS data with Medicare records makes the ICD-9-CM codes available for all 19,907 individuals from the 1999 NLTCS survey, including the genetic sample. Therefore, these data can be used to estimate the incidence rate of cognitive disorders for carriers and non-carriers of *APOE* e4.

Available empirical data on allele-specific incidence rates of cognitive disorders in the 1999 NLTCS *APOE* sample do not allow one to suggest any specific parametric function for these rates due to small sample sizes of genetic data. Therefore, we modeled incidence rates for carriers and non-carriers of the *APOE* allele e4 using four different models: (a) Gompertz's functions $\mu_i(x, 1999 - x) = a_i e^{b_i(x-65)}$ $i = 0, 1$ (0 stands for non-carriers of *APOE* allele e4); (b) linear functions $\mu_i(x, 1999 - x) = a_i + b_i(x - 65)$; (c) quadratic functions $\mu_i(x, 1999 - x) = a_i + b_i(x - 65) + c_i(x - 65)^2$; and (d) logistic functions

$$\mu_i(x, 1999 - x) = a_i e^{b_i(x-65)} \left/ (1 + \sigma_i^2 a_i b_i^{-1} (e^{b_i(x-65)} - 1)) \right.$$

Three sources of information were used in the estimation procedure: (a) data on the annual age-specific incidence of cognitive disorders for carriers and non-carriers of *APOE* allele e4; (b) data on age-specific proportions of carriers and non-carriers of the *APOE* allele e4 among individuals without cognitive disorders; and (c) annual data on the age-specific incidence of cognitive disorders occurring in the total 1999 NLTCS sample in years 2000 and 2001. We performed both the traditional analysis of genetic data maximizing the likelihood (1) and the joint analysis

of genetic and non-genetic data maximizing the likelihood (1), (2), (4), and estimated parameters of incidence rates for carriers and non-carriers of the *APOE* allele e4 modeled by different functions mentioned above. The results are presented in Fig. 3 and Table 2.

Statistical analysis

We compared the models with different hazard rates to evaluate which hazard rate gives the best fit to the data. As not all the models are nested, we compared the models' fit using the Akaike information criterion (Akaike 1974).

We tested the null hypotheses about equal hazard rates among carriers and non-carriers of *APOE* allele e4. For this purpose, we estimated models with equal parameters of hazard rates among carriers and non-carriers and general models with different hazard rates and performed the likelihood ratio test.

We calculated standardized incidence rates of cognitive disorders for carriers and non-carriers of *APOE* allele e4. The age distribution of the 1999 NLTCs participants (aged 65+) was used in the standardization procedure.

Software

All graphical and tabular information and statistical analyses presented in the paper were produced by self-written programs in MATLAB (Math Works 2004a). MATLAB's optimization toolbox (Math Works 2004b) was used to maximize the likelihood functions in the simulation study and in the applications to the NLTCs data. Respective computer codes are available upon request.

Results

Simulation study

To illustrate the accuracy of the estimates obtained in the analysis of data from a genetic sample alone, we calculated empirical distributions of parameter estimates from simulated data. The histograms of respective distributions are shown in Fig. 1 (gray columns). The mean values of each distribution are close to the real values of the respective parameter. However, the standard deviations of these distributions are large (Table 1). Thus, in the absence of additional information, the estimates of all parameters describing differences in genetic risks of disease development are unreliable.

Figure 1 also shows empirical distributions of the estimates of Gompertz's parameters obtained in the joint analysis of genetic data on hazard rates and genetic data on age trajectories of proportions, as well as non-genetic data on marginal hazard rates. One can see that the accuracy of the estimates is substantially improved: the respective standard deviations are substantially reduced (Table 1).

It is interesting to find out what sample size of genetic data is needed to obtain the accuracy of statistical estimates comparable to that reached in the joint analysis of genetic and non-genetic data. Figure 2 shows the results of analysis of simulated genetic data sets when information on allele/genotype-specific hazard rates is known for 20,000 individuals. One can see that the large sample size for genetic data substantially improves the accuracy of parameter estimates. The respective standard deviations are shown in Table 1. Comparison of distributions shown in Figs. 1 and 2 reveals that the use of additional information is equivalent to the substantial increase in the sample size of the data used in the direct genetic analysis (from 500 to 20,000 individuals).

Application to the NLTCS data

Table 2 summarizes the results of analysis of data on incidence rates of cognitive disorders in the 1999 NLTCS *APOE* sample using both the traditional and joint models. It shows estimates of parameters, standardized incidence rates for carriers and non-carriers of *APOE* allele $\epsilon 4$, and statistical comparison of models with different parametric specifications of incidence rates. The table shows (see column ‘AIC’) that the best fitting model in case of traditional models (estimating the genetic data set alone) is that with the Gompertz incidence rates, with the quadratic and logistic models providing a slightly worse fit, and the linear model providing the worst fit among all models. The last column shows *P*-values for the null hypothesis about equal incidence rates among carriers and non-carriers of *APOE* allele $\epsilon 4$. These results reveal that the application of the traditional models does not allow one to observe a statistically significant difference between the rates for carriers and non-carriers. The empirical analysis of allele-specific incidence rates does not permit reliable evaluation of these rates as well (see Fig. 3). The respective genetic sample sizes and numbers of newly detected cases of cognitive disorders are too small, especially for carriers of *APOE* allele $\epsilon 4$.

The last four rows of Table 2 show (see column ‘AIC’) that the best fitting model in case of joint models (estimating genetic and non-genetic data) is that with the logistic incidence rates, with the quadratic model providing the almost equivalent fit. The model with the Gompertz rates provides the worse fit in this case and the linear model provides the worst fit among all four models. However, Fig. 3 reveals that, in fact, three models (those with the logistic, quadratic and Gompertz rates) result in almost equivalent estimates of incidence rates (see Fig. 3a, b). The substantial difference in estimates of incidence rates for these models is observed only at extremely old ages (about 96+ years for non-carriers and about 88+ years for carriers) where the estimates become unreliable due to smaller sample sizes. In addition to this, the quadratic model also estimates a higher incidence rate for carriers at ages 65–66 years compared to the logistic and Gompertz models, which might be an artifact of the parametric specification of the rates as a quadratic function. The estimates of the ‘survival’ function (i.e., the probability of being free of cognitive disorders) are similar in the logistic, quadratic and Gompertz models (see Fig. 3c). The estimated proportions of non-carriers in these models are similar up to the oldest ages (about 90 years, see Fig. 3d). Thus, the application of joint models results in consistent estimates of incidence rates for carriers and non-carriers of *APOE* allele $\epsilon 4$ for the logistic, quadratic and Gompertz parameterizations of the rates. The last column in Table 2 shows that the likelihood ratio test applied to the joint models with the quadratic and Gompertz incidence rates reveals a statistically significant difference between incidence rates for carriers and non-carriers of *APOE* allele $\epsilon 4$.

The dynamics of age-specific incidence rates for carriers vs. non-carriers is similar in all four parameterizations: the risk of development of cognitive disorders among *APOE* $\epsilon 4$ carriers is higher—and increases faster—than among non-carriers of this allele. However, there are two exceptions: a decline in the quadratic rates and smaller Gompertz rates for carriers at ‘younger’ ages. Nevertheless, standardized incidence rates are consistently higher for carriers and the differences between standardized rates for carriers and non-carriers become higher in case of the joint model compared to the traditional one (for all parameterizations except the linear rates).

Discussion

Substantial improvement in the accuracy of statistical estimates resulting from the joint analysis of genetic and non-genetic data opens a new avenue in genetic studies of aging and aging-related disorders. The thing of practical importance is that additional non-genetic data used in the joint analysis are usually available and easily accessible. The application of this approach to the joint analysis of 1999 NLTCS genetic and non-genetic data allowed us to consistently

estimate the risk of development of cognitive disorders among *APOE* e4 carriers and non-carriers. In this study, the traditional empirical analysis of allele-specific incidence rates did not produce reliable results due to sample size limitations. Moreover, the joint analysis of genetic and non-genetic data also permitted rejecting the null hypothesis about equal incidence rates for carriers and non-carriers in cases where the application of the traditional modeling based on analyses of genetic data set alone did not allow us to reveal statistically significant differences.

The cross-sectional nature of genetic data may be a serious limitation for genetic analyses of morbidity and mortality in cases where the genetic structure of subsequent cohorts represented in the cross-sectional analysis is not similar. This could be the case when initial frequencies of respective alleles in such cohorts have substantial variation due to migration—or other reasons. This problem can also be addressed in the joint analysis of genetic and non-genetic data by assuming different initial frequencies in the cohorts comprising a cross-sectional sample. Table 3 summarizes the results of a simulation study estimating genetic and non-genetic data from a cross-sectional sample when the initial frequencies of selected allele in subsequent cohorts experience a linear trend. The first line in this table shows the mean values and the standard deviations of parameters of the Gompertz's curves for carriers and non-carriers of selected (hypothetical) allele evaluated in the analysis of genetic data set alone. One can see that parameter estimates are slightly biased and the standard errors are too large to make reliable conclusions about parameter values. The second line shows the mean values and the standard deviations of the estimates of Gompertz's parameters evaluated in the joint analysis of genetic and non-genetic data without taking into account the presence of a linear trend in the initial frequencies of the selected allele in subsequent cohorts. One can see that the mean values in such an analysis are biased. The third line shows the mean values and the standard deviations of the parameters estimates produced in the joint analysis of genetic and non-genetic data when the trend in the initial values of genetic frequencies is taken into account in the estimation algorithm. One can see that in this case the parameter estimates become unbiased and the standard deviations are substantially reduced compared to the analysis of genetic data alone.

Another important limitation involves secular trends in human morbidity and mortality observed during the last century. These trends take place because of changes in influential factors such as economic and living conditions, nutrition, life style, etc. Even if these changes developed in a relatively stable initial genetic background for all cohorts, they may still have resulted from different interactions between genes and environment that in turn may have different effects for carriers and non-carriers of a respective allele (genotype). Such limitations are, however, common for all genetic association studies (Yashin et al. 1999,2000; Lewis and Brunner 2004). To evaluate the sensitivity of these factors in results of genetic analysis, additional simulation studies are needed. Some of such studies have been performed in Yashin et al. (1999,2000) and Tan et al. (2002) in the analysis of genetic data obtained in centenarian studies. Note that the approach described above can be used in the evaluation of effects of non-genetic factors as well, under the condition that these factors did not experience changes over the individuals' life course.

Note that additional data, that can make an important contribution to the accuracy of statistical estimates in a joint analysis with genetic samples, are not restricted to those described and used in this paper. Epidemiological data on the effects of genetic or non-genetic risk factors evaluated for selected age groups and data on secular trends in human mortality and morbidity by cause also contain hidden information on respective genetic influence. Methods for a joint analysis of such data still have to be developed. It would be also interesting and important to explore the opportunity of using other approaches (Devlin et al. 2001; Pritchard et al. 2000) in the joint analysis of data from different sources to reduce the effects of population stratification.

Sources of additional errors in statistical estimates of genetic parameters are discussed in Schork (2002).

Acknowledgements

This research was supported by NIH/NIA grants PO1 AG08761-01, U01 AG023712 and 5U01-AG-007198-18. The authors thank V. Lewis for help in preparing this manuscript for publication.

References

- Akaike H. A new look at the statistical model identification. *IEEE Trans Automatic Control* AC-19 1974;716–723.
- Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001;2:91–99. [PubMed: 11253062]
- Carrieri G, Bonafè M, De Luca M, Rose G, Varcasia O, Bruni A, Maletta R, Nacmias B, Sorbi S, Corsonello F, Feraco E, Andreev KF, Yashin AI, Franceschi C, De Benedictis G. Mitochondrial DNA haplogroups and APOE4 allele are non-independent variables in sporadic Alzheimer's disease. *Hum Genet* 2001;108:194–198. [PubMed: 11354629]
- De Benedictis G, Tan Q, Jeune B, Christensen K, Ukraintseva SV, Bonafe M, Franceschi C, Vaupel JW, Yashin AI. Recent advances in human gene-longevity association studies. *Mech Ageing Dev* 2001;122:909–920. [PubMed: 11348658]
- Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;60:155–166. [PubMed: 11855950]
- Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Statist* 2001;55:19–24.
- Lewis SJ, Brunner EJ. Methodological problems in genetic association studies of longevity – the apolipoprotein E gene as an example. *Int J Epidemiol* 2004;33:962–970. [PubMed: 15319409]
- Math Works. MATLAB: the language of technical computing. Programming. Version 7. The Math Works, Inc; Natick, MA: 2004a.
- Math Works. User's guide. Version 3. The Math Works, Inc; Natick, MA: 2004b. Optimization toolbox for use with MATLAB.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000;67:170–181. [PubMed: 10827107]
- Schork NJ. Power calculations for genetic association studies using estimated probability distributions. *Am J Hum Genet* 2002;70:1480–1489. [PubMed: 11992254]
- Tan Q, De Benedictis G, Ukraintseva SV, Franceschi C, Vaupel JW, Yashin AI. A centenarian-only approach for assessing gene-gene interaction in human longevity. *Eur J Hum Genet* 2002;10:119–124. [PubMed: 11938442]
- Todd JA. Tackling common disease. *Nature* 2001;411:537–539. [PubMed: 11385552]
- Varcasia O, Garasto S, Rizza T, Andersen-Ranberg K, Jeune B, Bathum L, Andreev K, Tan Q, Yashin AI, Bonafe M, Franceschi C, De Benedictis G. Replication studies in longevity: puzzling findings in Danish centenarians at the 3'APOB-VNTR locus. *Ann Hum Genet* 2001;65:371–376. [PubMed: 11592926]
- Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L, Franceschi C. Genes, demography, and life span: the contribution of demographic data in genetic studies on aging and longevity. *Am J Hum Genet* 1999;65:1178–1193. [PubMed: 10486337]
- Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, Valensin S, De Luca M, Carotenuto L, Franceschi C. Genes and longevity: lessons from studies of centenarians. *J Gerontol A Biol Sci Med Sci* 2000;55A:B319–B328. [PubMed: 10898245]

Abbreviations

AD

Alzheimer's disease

AIC

the Akaike information criterion

APOE

apolipoprotein E

ICD-9-CM

International Classification of Diseases, 9th Revision, Clinical Modification

NLTCS

the National Long Term Care Survey

Fig 1.

Simulation study: a substantial improvement in the accuracy of estimates in the joint analysis of genetic and non-genetic data. The figure shows distributions of estimates of Gompertz's parameters describing hazard rates for non-carriers, a_0 (a) and b_0 (b), and carriers, a_1 (c) and b_1 (d), of hypothetical allele obtained from 1,000 simulated data sets estimated using genetic data alone ('Genetic sample') and using the joint analysis of genetic and non-genetic data ('Joint data'). The symbols • show the locations of true parameter values

Fig 2.

Simulation study: a comparable improvement in the accuracy of estimates reached by the substantial increase in the sample size of genetic data. The figure shows the distributions of estimates of Gompertz's parameters describing hazard rates for non-carriers, a_0 (a) and b_0 (b), and carriers, a_1 (c) and b_1 (d), of hypothetical allele obtained from 1,000 simulated data sets estimated using genetic data alone in case of different sample sizes: 500 and 20,000 individuals. The symbols • show the locations of true parameter values

Fig 3.

The results of joint analysis of genetic and non-genetic data: application to the 1999 NLTCS genetic sample. **(a)** Empirical estimates of incidence rates of cognitive disorders for non-carriers of *APOE* allele e4 in the genetic sample ('*APOE* sample') and the estimates obtained from the joint analysis of genetic and non-genetic data using different parameterizations of incidence rates ('Gompertz,' 'linear,' 'logistic,' 'quadratic'). **(b)** Empirical estimates of incidence rates of cognitive disorders for carriers of *APOE* allele e4 in the genetic sample ('*APOE* sample') and the estimates obtained from the joint analysis of genetic and non-genetic data. **(c)** Probability of staying in the state of normal cognitive functioning ('survival' function) calculated from the mixture of carriers and non-carriers of *APOE* allele e4 using: (1) the 1999 NLTCS genetic sub-sample ('*APOE* sample'); (2) the 1999 NLTCS data ('NLTCS sample'); and (3) the joint analysis of genetic and non-genetic data. **(d)** Estimates of the age trajectories of proportions of non-carriers of *APOE* allele e4 among individuals without cognitive disorders obtained from the 1999 NLTCS genetic sample ('*APOE* sample') and from the joint analysis of genetic and non-genetic data

Table 1
Simulation study: mean values (standard deviations) of parameter estimates for 1,000 data sets in different models

Mixture	Genetic sample		Sample sizes		Parameters				
	Proportions	Hazard	Genetic sample	Mixture	p	a_0	b_0	a_1	b_1
No	No	Yes	500	–		0.008 (0.001)	0.08 (0.005)	0.0049 (0.0013)	0.13 (0.011)
Yes	Yes	Yes	500	20,000	0.75 (0.0008)	0.008 (0.0002)	0.08 (0.0008)	0.005 (0.0002)	0.13 (0.002)
No	No	Yes	20,000	–		0.008 (0.0002)	0.08 (0.0007)	0.005 (0.0002)	0.13 (0.002)
True values:					0.75	0.008	0.08	0.005	0.13

Table 2

Application of different models to the 1999 NLTCS *APOE* sample: estimates of parameters, standardized incidence rates for carriers and non-carriers of *APOE* allele ϵ_4 , and statistical comparison of models

Data (model)	Incidence	Parameters			Standardized rates			AIC ^a	P-values ^b
		<i>p</i>	<i>a</i> ₀	<i>b</i> ₀	$\epsilon_0(\sigma_0)$	<i>a</i> ₁	<i>b</i> ₁	$\epsilon_1(\sigma_1)$	
Genetic data (traditional model)	Gompertz		0.0056	0.0892		0.0082	0.0912		0.494
	Linear		0.0014	0.0016		0.0101	0.0018		0.468
	Logistic		0.0056	0.0896	0.00062	0.0082	0.0910	0.01089	0.703
	Quadratic		0.0153	-0.0015	0.00012	0.0541	-0.0086	0.00042	0.458
Genetic and non-genetic Data (joint model)	Gompertz	0.756	0.0079	0.0800		0.0053	0.1285		0.002
	Linear	0.756	0.0000	0.0022		0.0000	0.0029		0.189
	Logistic	0.755	0.0049	0.1314	1.00163	0.0058	0.1314	0.56222	0.127
	Quadratic	0.752	0.0000	0.0014	0.00005	0.0272	-0.0049	0.00035	0.031

^a Difference between values of the Akaike information criterion (AIC) for respective models and the best fitting model with minimal AIC (separately for the traditional and joint models)

^b P-values for the likelihood ratio test applied to respective models and models with equal parameters of incidence rates among carriers and non-carriers (the null hypothesis: incidence rates among carriers and non-carriers are equal)

Table 3

Simulation study: mean values (standard deviations) of parameter estimates for different models estimating 1,000 data sets with a linear over-time trend in the initial proportion (p) of allele/genotype

Trend in p is estimated	Genetic sample		Mixture		Sample sizes		Parameters								
	Proportions		Hazard		Marginal hazard		Genetic sample		Mixture		$p_0(p)$				
											a_0	b_0	a_1	b_1	q_0
No	No	Yes	No	Yes	No	500	–				0.009 (0.0058)	0.08 (0.027)	0.007 (0.0086)	0.14 (0.058)	
No	Yes	Yes	Yes	Yes	Yes	500	20,000				0.01 (0.0009)	0.07 (0.004)	0.003 (0.0005)	0.15 (0.007)	
Yes	Yes	Yes	Yes	Yes	Yes	500	20,000				0.75 (0.007)	0.08 (0.0013)	0.005 (0.0016)	0.13 (0.011)	–0.0024 (0.0014)
							True values:				0.75	0.08	0.005	0.13	–0.0025