Research Project sponsored by the Society of Actuaries Committee
on Life Insurance Research (CLIR)

Final Report

Living to 100 and Beyond:

Search for Predictors of Exceptional Human Longevity

Natalia S. Gavrilova, Leonid A. Gavrilov

Center on Aging, NORC at the University of Chicago

October 4, 2005

Address for correspondence:
Dr. Natalia S. Gavrilova, Center on Aging
NORC/University of Chicago
1155 East 60th Street, Chicago, IL 60637
Fax: (773) 256-6313; Phone: (773) 256-6359
E-mail: gavrilova@longevity-science.org

Abstract

Centenarians (people living to age 100 and beyond) represent one of the fastest-growing age groups of the American population, with obvious implications for actuarial science and practice. Yet, factors predicting exceptional longevity and its time trends remain to be fully understood. In this project we explored the new opportunities provided by the ongoing revolution in information technology, computer science and Internet expansion for studies of exceptional human longevity. Specifically, we explored the availability and quality of computerized online genealogies of long-lived individuals by cross-checking them with other Internet resources, including the Social Security Administration (SSA) Death Master File (DMF) and the early U.S. censuses. To this aim, we extracted detailed family data for 991 alleged centenarians born in 1875-1899 in the United States from publicly available computerized genealogies of 75 million individuals identified in our previous study (Gavrilova, Gavrilov, 1999). In order to validate the age of the centenarians, we linked these records to the Social Security Administration Death Master File records (for death date validation) and then to the records of the U.S. censuses for years 1900, 1910 and 1920 (for birth date validation). The results of this cross-validation study demonstrated that computerized genealogies might serve as a useful starting point for developing a family-linked scientific database on exceptional human longevity.

This report also presents some preliminary studies on predictors of exceptional human longevity, including familial factors and early-life living conditions. Specifically, this study suggests that there may be a link between exceptional longevity and a person's birth order. We found that first-born daughters are three times more likely to survive to age 100, compared to later-born daughters of higher birth orders (7+). First-born sons are twice more likely to become centenarians compared to sons having birth order between four and six. There is also a profound

sex difference in the effects of birth order on human longevity. For sons, this dependence has an unusual U-shaped form, with highest longevity chances for both the first-born and the last-born (9+) sons.

The comparison of households where children –(future centenarians) were raised (using data obtained through linkage of genealogies to early U.S. censuses) with control households drawn from the Integrated Public Use Microdata Series (IPUMS) for the 1900 U.S. census suggests that a farm background (farm ownership by parents in particular) and child residence in the Western region in the United States may be predictive for subsequent survival to age 100. These findings are consistent with the hypothesis that lower burden of sickness during childhood (expressed as lower child mortality in families of farm owners and families living in the West) may have far-reaching consequences for survival to extreme old ages.

Data from the Social Security Administration Death Master File allowed us to analyze mortality patterns at advanced ages, using the method of extinct generations. The study of four birth cohorts (1885, 1886, 1889 and 1891) found that mortality grows steadily with age from 90 to 105 years with almost no obvious signs of expected mortality deceleration. After age 105, the mortality estimates become less reliable because of significant statistical noise. The mortality of males exceeds the mortality of females up to very old ages (110 years). It also was found that life expectancy at age 80 depends on the month of a person's birth: individuals born in January live longer lives than persons born in other months and in April-June in particular. This periodicity repeats in every studied birth cohort starting from birth year 1885 to 1899. However, by age 100 this dependence of survival on month of birth fades out, indicating that centenarians indeed represent a selected population.

TABLE OF CONTENTS

LIST OF FIGURES

Figure 6. Hazard rate (mortality force, year^{-1}) for 1891 birth cohort. Less reliable data for Southern states, Puerto Rico and Hawaii are excluded. Data from the Social Security Administration Death Master File.

Figure 7. Hazard rate (mortality force, year^{-1}) for males and females from 1886 birth cohort. Data are fitted using quadratic polynomial regression. Total U.S. population.

Figure 8. Hazard rate (mortality force, year^{-1}) for males and females from 1886 birth cohort. Lifespan is estimated in whole years. Data are fitted using quadratic polynomial regression. Total U.S. population.

Figure 9. The dependence of life expectancy at age 80 on person's month of birth. Comparison of 1885 and 1891 birth cohorts. Data on extinct birth cohorts obtained from the Social Security Administration Death Master File. Total U.S. population.

Figure 10. Periodic seasonal changes in life expectancy at age 80 for 1885-1899 birth cohorts depending on month of birth. Total U.S. population.

Figure 11. Lack of seasonal dependence of life expectancy at age 100 on person's month of birth. Comparison of 1885 and 1891 birth cohorts. Total U.S. population.

Figure 12. Life expectancy at age 100 for 1885-1893 birth cohorts. Total U.S. population.

LIST OF TABLES

I. INTRODUCTION

Centenarians (people living to age 100 and beyond) represent one of the fastest-growing age groups of the American population, with obvious implications for actuarial science and practice. The number of U.S. centenarians is growing at a rate of about 4.1 percent per year, so their numbers have increased by 51 percent in the 10-year period from Jan. 1, 1990 to Jan. 1, 2000 (Kestenbaum, Ferguson, 2005). Yet, factors predicting exceptional longevity and its time trends remain to be fully understood. Thorough and comprehensive studies of survival at advanced ages requires searching for new data sources in addition to careful re-evaluation of already known ones.

Our previous search for additional data resources (see Gavrilov, Gavrilova, 1998; Gavrilova, Gavrilov, 1999) revealed an enormous amount of new family lifespan data that could be made readily available for subsequent full-scale studies. Millions of genealogical records are already computerized and could be used for the study of familial and other predictors of human longevity (after strict data validation). Most of these genealogies are a product of family reconstitution, carried out both by professional genealogists and by family members tracing their ancestry back to the founder who brought their surname to America or even to their European family roots. The compilers of genealogies aided this time-consuming task by using many different sources: genealogical libraries, The Church of Jesus Christ Latter-day Saints (Mormon) family history centers, genealogical search engines available on the Internet, computer CDs with census, marriage, land and probate records, and many other resources for genealogical research.

Computerized genealogies provide the most complete information on the lifespan of centenarians' relatives, compared to other data sources, such as death certificates, census data and the Medicare database. Census records provide information on birth years of parents and siblings, but no information on death dates is available. The Medicare database allows identification of spouses (see Iwashyna et al., 1998), but no information on parents and other relatives is available. The Social Security Administration NUMIDENT file contains information on the names of parent/child pairs for Medicare beneficiaries (65 years and older). In the latter case, however, one cannot obtain information on distant ancestors (e.g., grandparents) as well as other relatives (e.g., first cousins), so there is no opportunity for reconstruction of pedigrees. Thus, computerized genealogies provide unique opportunities to greatly expand the scope of actuarial studies of human longevity, and they cannot be adequately substituted by other data resources.

The main purpose of this exploratory project is to identify and validate new data resources for actuarial studies of predictors of exceptional human longevity, and to ensure feasibility of a subsequent large-scale research program on related topics (by developing unique datasets, new computer techniques and new substantive hypotheses on determinants of human longevity supported by preliminary studies of this project).

The first goal of this project is to develop a high-quality, family-linked database on American centenarians, rich in the scope of individual predictor variables (including both early-life and life-course characteristics as well as a familial history of longevity), and complete for each predictor variable (without missing observations).

In this report we describe our experience in the identification, collection, verification and analysis of data taken from computerized genealogies for long-lived individuals. The process of data quality evaluation and centenarians' age verification is described in detail because it appears to be the first attempt in systematic assessment of quality for this new and potentially promising data source on family factors of longevity. We also test several hypotheses on the effects of early-life living conditions and family factors affecting lifespan.

The idea of fetal origins of adult degenerative diseases and early-life programming of late-life health and survival is being actively discussed in the scientific literature (Lucas, 1991; Gavrilov, Gavrilova, 1991; 2003a; Barker, 1998; Kuh, Ben-Shlomo, 1997; Lucas *et al*., 1999; Costa, Lahey, 2003). The historical improvement in early-life conditions may be responsible for the observed significant increase in human longevity through the process called "technophysio evolution" (Fogel, Costa, 1997). Additional arguments suggesting the importance of early-life conditions in later-life health outcomes are coming from the reliability theory of aging and longevity (Gavrilov, Gavrilova, 1991; 2001a; 2003a). According to this theory, biological species (including humans) are starting their lives with an extremely high initial load of damage, and, therefore, they should be sensitive to early-life living conditions affecting the level of initial damage (Gavrilov, Gavrilova, 1991; 2001a; 2004).

The concept of high initial damage load also predicts that early life events may affect survival in later adult life through the level of initial damage. This prediction proved to be correct for such early-life indicators as the parental age at a person's conception (Gavrilov, Gavrilova, 1997, 2000, 2003b; Gavrilova et al., 2003) and the month of person's birth (Gavrilov, Gavrilova, 1999, 2003b, Gavrilova et al., 2003; Doblhammer, 1999; Doblhammer, Vaupel, 2001; Costa, Lahey, 2003). There is mounting evidence now in support of the idea of fetal origins of adult degenerative diseases (Barker, 1998; Kuh, Ben-Shlomo, 1997; Lucas, Fewtrell, Cole, 1999), and early-life programming of aging and longevity (Gavrilov, Gavrilova, 1991, 2001, 2003a; Gavrilova et al., 2003). We tested some of these ideas in the presented study, as you will see later on.

In this project we also explored the feasibility of using data from the Social Security Administration Death Master File (SSA DMF) for detailed estimates of mortality rates at advanced ages and studies of month-of-birth effects on human longevity.

It is now considered an established fact that mortality at advanced ages has a tendency to deviate from the Gompertz law, so the logistic model is often used to fit mortality (Perks, 1932; Beard, 1959; 1971; Horiuchi, Wilmoth, 1998). It was Gompertz (1825) himself who first noted this phenomenon of late-life mortality deceleration, and later other actuaries had suggested a logistic formula in order to account for mortality deceleration at advanced ages (Perks, 1932; Beard, 1959; 1971). Greenwood and Irwin (1939) provided a detailed description of this phenomenon in humans and even made the first estimates for the asymptotic upper limit of human mortality. According to their estimates, the mortality kinetics of long-lived individuals is close to the law of radioactive decay with the half-life period approximately equal to one year.

The same phenomenon of "almost non-aging" survival dynamics at extreme old ages is detected in many other biological species (Sacher, 1966; Economos, 1979; 1980; 1983; 1985; Curtsinger et al., 1992; Carey et al., 1992; Vaupel et al., 1998). In some insects, mortality plateau can occupy a sizable part of their life (Carey et al., 1992). The existence of mortality plateaus is well established for a number of lower organisms, mostly insects. In the case of mammals, data are much more controversial. Although Lindop (1961) and Sacher (1966) reported short-term periods of mortality deceleration in mice at advanced ages, and even used Perks formula in their analyses, Austad (2001) recently argued that rodents do not demonstrate mortality deceleration, even in the case of large samples. Study of baboons found no mortality deceleration at advanced ages (Bronikowski et al., 2002). In the case of humans, this problem is not yet resolved completely. Data for extremely long-lived individuals are scarce and subjected to age exaggeration. Traditional demographic methods of mortality analysis based on period life tables suffer from the well-known population denominator problem. More accurate estimates of mortality at advanced ages can be obtained using the method of extinct generations (Vincent, 1951; Depoid, 1973). In the Kannisto-Thatcher database (Thatcher, 1999) mortality is estimated by the method of extinct generations and data are aggregated for several calendar periods in order to accumulate enough cases of survivors to older ages. The aggregation of deaths for several calendar periods, however, creates a heterogeneous mixture of cases from different birth cohorts. Mortality deceleration observed in these data might be a result of data heterogeneity (Beard, 1959; 1971). In addition to that, many assumptions about distribution of deaths in the age/time interval used in mortality estimation are not valid for extreme old ages when mortality is particularly high. Thus, we need more research efforts to obtain reliable estimates of mortality at advanced ages. In this project we analyzed mortality at advanced ages using data for large U.S. extinct birth cohorts provided by the SSA DMF.

The SSA DMF data allow us to explore another interesting problem: the effects of early-life conditions and month-of-birth in particular on mortality at advanced ages. It was shown that month of birth has a significant effect on later-life mortality and lifespan (Gavrilov, Gavrilova, 1999; Doblhammer, Vaupel, 2001; Costa, Lahey, 2003). For example, Costa and Lahey (2003) used data on month of birth and mortality for the Union Army veterans at age 60-79 in 1900 and Americans of the same age in 1960-1980. They found that persons born in the second quarter had higher mortality than persons born in the fourth quarter (Costa, Lahey, 2003). Another study of month-of-birth effects on mortality in the United States (Doblhammer, 2003) computed the mean age at death using cross-sectional death certificates that do not take into account the underlying structure of populations exposed to risk. This approach could be justified only for stationary populations with population age structures constant over time. In real life this assumption usually is not valid, and the mean age at death (calculated from death certificates) is affected by temporal trends in population characteristics as well as temporal changes in seasonality of births and infant mortality. Because of these problems, using mean age at death from cross-sectional rather than cohort data as a proxy for life expectancy may lead to incorrect conclusions about better survival for low-educated or widowed persons (as follows from the data published by Doblhammer, 2003). More reliable estimates of mortality by month of birth could be obtained either by using death certificates in conjunction with population denominator data taken from censuses, or by analyzing cohort mortality. In this regard the SSA DMF containing cohort data may provide more accurate estimates of month-of-birth effects on mortality, and we applied this data resource in the current study.

The results presented in this report demonstrate that actuarial studies on human longevity could be modernized and advanced further by using new computerized data resources and new research ideas on longevity predictors, as well as some new findings described in this report.

II. MATERIALS AND METHODS

<u>Survey of the existing computerized genealogies</u>

At the first stage of the project, we made a survey of the relevant data resources and identified computerized family histories for over 75 million deceased individuals using collections of online genealogies (see Appendix I for details) identified in our previous studies (Gavrilov, Gavrilova, 1998; Gavrilova, Gavrilov, 1999). Centenarian family histories were drawn from computerized family trees using the following selection criteria: (a) persons should have both the birth and death date information and have lifespans of 100 years and over, (b) persons should be born in the United States after 1875 and (c) persons should have pedigree information for at least three generations of ancestry (on both the paternal and maternal sides) as well as information on birth date and death date of parents.

The decision to exclude foreign-born centenarians from our study was conditioned by the difficulties of their age verification. The main obstacle here is that for persons born in the study window of 1890-1900 it may be difficult to find many foreign-born persons in the available early U.S. censuses (1900 and 1910) used for birth date verification, because many of these persons could immigrate later. Also, in the case of foreign-born persons, the U.S. census data are useless in providing information about early-life conditions because foreign-born children spent a part of their childhood abroad in unknown living conditions. Therefore, because this particular project focused on the role of early-life conditions in predicting exceptional human longevity, foreign-born persons are less informative than U.S.-born persons. In addition, it is particularly difficult to verify the quality of genealogical data for foreign-born centenarians. Thus, by excluding the genealogies for foreign-born centenarians, we excluded the most questionable part of the data, which are particularly difficult to cross-validate through early U.S. censuses. It should also be noted that foreign-born children comprised a small proportion (3 percent) of all children below age 10 enumerated in the 1900 census. We obtained this estimate from the Integrated Public Use Microdata Series (IPUMS) 1 percent random sample of the 1900 U.S. census population (For more details on the IPUMS project see Ruggles et al., 2004).

Using online genealogical data resources, we identified over 2,000 genealogies, which contained detailed information about long-lived persons as well as detailed information about their parents and grandparents. The obtained genealogies were recorded in the so-called

GEDCOM data format, which is used for genealogical data exchange (Gavrilova, Gavrilov, 1999) and is described below.

Description of the GEDCOM format

Although all genealogical software has its own data format, the genealogical data are shared among other genealogists through the GEDCOM format. "GEDCOM" stands for the "GEnealogical Data COMmunication" standard proposed by the Family History Department of The Church of Jesus Christ of Latter-day Saints (LDS Church) and adopted by many developers and users of genealogical software (Family History Department, 1996). The purpose of GEDCOM is to simplify the exchange of computerized historical and genealogical information. GEDCOM files are created in ASCII (text) format with special tags at the beginning of each line related to specific family information (variables). The most common variables contain personal information (name, birth date and place, death date and place) and family information (links to spouses and children and links to parents and siblings). In many cases, GEDCOM files contain more detailed information (occupation, education, residence, title, religion, cause of death, burial place and special notes). Data on living individuals are eliminated in the majority of computerized genealogies (to protect their privacy) except for their names and family links.

Information contained in GEDCOM files cannot be immediately used in statistical analyses because it needs to be converted to a relational database for data cleaning and further analysis. Thus, after collecting data in the form of GEDCOM files, they were converted to the relational database (MySQL) for their further verification and analysis.

Database on U.S. centenarians

We used the Entity-Relationship (ER) approach to database modeling (Bagui and Earp, 2003). The data model focuses on what data should be stored in the database. To put this in the context of a relational database, the data model is used to design the relational tables. To do this, we first created an entity-relationship diagram for our model, which represents the data structures in pictorial form. Genealogical data can be well represented by a two-entity design: persons and unions/marriages with one-to-many relationships (one person may have many unions while a particular union/marriage describes a unique pair of partners). This design was further extended by adding an entity reflecting the SSA DMF data and two entities for the early census data: households and household members. Physical realization of this model was made using a common set of software: Apache Web server, PHP program language for user interface and the MySQL database management system.

The collected GEDCOM files for centenarians were screened for long-lived individuals and converted to the MySQL database using specially developed program scripts. As a result, we obtained information for 2,004 long-lived individuals and their relatives (including parents and grandparents) in the form of a relational database. From these 2,004 records for long-lived individuals, we selected 991 records for centenarians born in the United States after 1875.

As with any new data resource, this data set has an uncertain quality, which requires additional efforts for data verification and quality control using several independent data sources. Our primary concern was the possibility of incorrect dates reported in genealogies. Previous studies found that age misreporting and age exaggeration in particular are more common among long-lived individuals (Hill et al., 2000; Rosenwaike, Stone, 2003; Shrestha, Preston, 1995). For this reason the focus of our study was on age verification for long-lived individuals rather than for other members of genealogy.

Verification of centenarian birth and death dates

Verification of centenarian birth and death dates was made in three stages.

Data consistency checks

To verify the centenarian's birth date, we first compared the person's birth date with birth dates for the person's parents, as well as with birth and marriage dates for the person's spouses (data consistency test). This was the preliminary test followed by two more sophisticated tests for data quality as described later.

Validation of death dates for U.S. centenarians using the Social Security Administration Death Master File

In this project we followed the approach of age verification and data linkage developed by demographers at the University of Pennsylvania (Rosenwaike, Logue, 1983; Preston et al., 1996; Rosenwaike et al., 1998; Hill et al., 2000; Rosenwaike, Stone, 2003).

The verification of death dates is an important step in quality control because it eliminates cases with potential mistakes and misprints in death dates reported for alleged centenarians. The verification of death dates was accomplished through a linkage of genealogical data to the SSA DMF. This is a publicly available data source that allows a search for individuals using various criteria: birth date, death date, first and last names, Social Security number and place of last residence (see Appendix I). This resource covers deaths that occurred in the period 1937-2003 (see Faig, 2001, for more details). Many researchers suggest that the quality of SSA/Medicare data for older persons is superior to vital statistics records because of strict evidentiary requirements in applying for Medicare, while age reporting in death certificates is made by proxy informant (Kestenbaum, 1992; Kestenbaum, Ferguson, 2001; Rosenwaike et al., 1998; Rosenwaike, Stone, 2003). In this project we based the death date verification on linkage to the SSA DMF, which is publicly available at the Rootsweb Web site (Faig, 2001).

<u>Verification of centenarian birth dates using the early U.S. censuses</u>

In order to verify centenarian birth dates, data for centenarians were checked against the early U.S. census records collected when the centenarian was a child or young adult. For validation purposes, the early U.S. censuses (1900, 1910 and 1920) are particularly important, because they provide information on future centenarians during their childhood and early adulthood years when age exaggeration is less common compared to claims of exceptional longevity made at old age. The preference was given to the 1900 census because it is more complete and detailed (in regard to age verification) compared to the 1910 and 1920 censuses. Specifically, the 1900 U.S. census provided year and month of birth, not just an age at enumeration date.

Information from the 1900 U.S. census was used not only for age verification, but also for a study of early-life factors and chances to survive to extreme old ages. The 1900 U.S. census provides the following information for household and its members: state, county, and township of residence; street and house number (where available); relationship to head-of-household; gender and ethnicity; month and year of birth and age at last birthday; marital status and, if married, length of marriage; for married women, number of children born and number living; birthplace of person and birthplaces of mother and father; for aliens or naturalized citizens, year of immigration and citizenship status; occupation of each person 10+ years and number of months not employed; information about school attendance and literacy; and information about home ownership or farm residence. An important advantage of the 1900 U.S. census is the availability of information about year and month of birth, providing an additional source for birth date verification.

In our study, the linkage of centenarian records to the early census data is facilitated by online availability of the entire indexed U.S. 1900, 1910 and 1920 censuses—a service provided by Genealogy.com and Ancestry.com (see Appendix I). In our project we conducted a linkage of 534 centenarian records (for centenarians found in the SSA DMF with confirmed centenarian status and born after 1889) to the early U.S. censuses. If individuals were not found in the 1900 census, then attempts were made to locate them in the 1910, 1920 and 1930 censuses.

Study of possible links between birth order and exceptional longevity

Information about birth order of centenarians allowed us to test a hypothesis questioning whether the centenarians are distributed randomly within a sibship (brothers and sisters in the family) or not. If a centenarian's birth order is determined by chance only and is not linked to exceptional longevity, then the ratio of [*centenarian birth order/(family size + 1)*] should be equal to 0.5 on average. If centenarians are found more often among the older or among the younger siblings, then the observed ratio, named "centenarian birth order ratio" (CBOR), should demonstrate a statistically significant deviation from the expected value of 0.5.

In order to test a hypothesis questioning whether centenarians are distributed randomly within a sibship we also used another statistic named 'centenarian birth order difference' (CBOD): [*centenarian birth order - (family size + 1)/2*]. If centenarians are distributed randomly by birth order within a sibship (independently of their centenarian status), then this difference should be equal to zero on average.

To study the birth order effects in more depth and to see how the odds of survival to age 100 depend on birth order we have applied a logistic regression model with a binary outcome variable (becoming a centenarian or not) and with two predictor variables (birth order and family size) included in a polynomial-fitting model with non-centenarian siblings born in the same time window used as a control group.

Study of early-life predictors of longevity

The resulting dataset of 1900 and 1910 households linked to centenarian genealogies allows us to make a comparison of these households to the general set of households enumerated in early censuses. We followed in part the methodological lines established by Preston et al., 1998, and used individual data from the 1900 U.S. census of population as a control group. The data are available as part of the IPUMS from the University of Minnesota (Ruggles et al., 2004). The sample represents 1 percent of Caucasian households enumerated in 1900. The linkage to early U.S. censuses in our study found that most centenarians in our sample were Caucasian (with the exception of two American Indian families), so we used a sample of the Caucasian population from the IPUMS as a control. At this initial stage of data analysis, we conducted a comparison of households that raised a future centenarian (linked to the 1900 census) to the general sample of Caucasian households enumerated by the 1900 census, which had children below age 10 (to make these households comparable to our set of centenarians who were born in 1890-1899 and hence were below age 10 in 1900).

We applied a method of multiple logistic regression (procedure 'logistic' in the Stata statistical package) in order to compare the two sets of households. The set of variables describing a household is similar to one applied by Preston et al. (1998). We did not use the variable describing the occupation of the father because this variable is strongly correlated with ownership and farm status variables and because of existing problems with the classification of diverse occupations. In fact, 63 percent of the fathers of centenarians in our sample were farmers by occupation, almost all white-collar fathers owned their house, and most low-skilled fathers were renters.

Study of mortality using the SSA Death Master File

The collection of records from the SSA DMF undertaken in this validation study had some interesting ramifications for mortality analyses at advanced ages. The DMF collects deaths for persons who receive SSA benefits and currently covers over 90 percent of deaths occurring in the United States (Faig, 2002) and 93 percent to 96 percent of deaths of individuals aged 65 or older (Hill, Rosenwaike, 2001). Despite certain limitations, this data source allows researchers to obtain detailed estimates of mortality at advanced ages.

In this study we collected information from the DMF available at Rootsweb.com on persons who lived 80 years and over and died before 2004. The total number of records collected is 9,014,591 including 924,222 records for persons who lived 100 years and over. The information contained in this file is interesting not only for verification purposes but also for mortality estimates at advanced ages. Several birth cohorts (those born in 1882-1891) may be considered extinct or almost extinct, so it is possible to apply the method of extinct generations (Vincent, 1951) and estimate mortality kinetics at very advanced ages up to 115-120 years. The DMF database is unique in this regard because it represents mortality experience for one of the largest cohorts of the oldest-old persons, which is readily available for survival analysis. Although the National Center for Health Statistics' National Death Index (NDI) provides superior coverage of deaths, its use is restricted and expensive, so for many researchers the DMF may be an appropriate choice (Hill, Rosenwaike, 2001).

The last deaths in the DMF available at the Rootsweb Web site occurred in January 2004 (when the present study was conducted). We obtained data for persons who died before 2004, because only two individuals born in 1885-1891 (birth cohorts that we studied) died in 2004. Thus, the 1885-1891 birth cohorts in this sample may be considered extinct or almost extinct. Assuming that the number of living persons belonging to these birth cohorts in 2004 is close to zero, it is possible to construct a cohort life table using the method of extinct generations, which was suggested and explained by Vincent (1951) and developed further by Kannisto (1994). In the first stage of our analyses we calculated an individual life span in completed months:

Lifespan in months = (death year – birth year) x 12 + death month – birth month

Then it is possible to estimate the hazard rate at each month of age using standard methods of survival analysis. All calculations were done using the Stata statistical package, procedures 'stset' and 'sts' (Stata Corp, 2005). This software provides nonparametric estimates of major survival functions including the Nelson-Aalen estimates of hazard rate (force of mortality). Note that a hazard rate, in contrast to a probability of death, q(x), has a dimension of time frequency, because of the time interval in the denominator (reciprocal time, time^{-1}). Thus the values of hazard rates depend on the chosen units of time measurement (day^{-1}, month^{-1} or year^{-1}). In this study survival times were measured in months, so the estimates of hazard rates initially had a dimension of month^{-1}. For the purpose of comparability with other published studies, which typically use the year^{-1} time scale, we transformed the monthly hazard rates to the more conventional units of year^{-1}, by multiplying these estimates by a factor of 12 (one month in the denominator of hazard rate formula is equal to 1/12 year). Also note that a hazard rate, in contrast to a probability of death can be greater than 1, and therefore its logarithm can be greater than 0 (and we indeed observed this at extreme old ages in some rare cases as will be described later). We estimated hazard rates for four single-year birth cohorts—those born in 1885, 1886, 1889 and 1891.

The SSA DMF does not provide information about the sex of the deceased. To avoid this limitation of the data sample, we conducted a procedure of sex identification using information about the 1,000 most commonly used baby names in the 1900s provided by the Social Security Administration (http://www.ssa.gov/OACT/babynames). These data come from a sample of 5 percent of all Social Security cards issued to individuals who were born during 1900s in the United States. From the lists of male and female names we removed names consisting of initials and names in which the sex was unclear (like Jessie or Lonnie). It is interesting to note that the SSA male list contains some obviously female names (Mary, Elizabeth) and the same problem was observed for the female list, which indicates that the SSA data apparently contain many sex misidentifications. These female names were removed from the male list and the same procedure was done for the female list. Using the final lists of male and female first names we identified the sex in 89.5 percent of cases of the 1886 birth cohort of persons aged 90 years and over. The remaining 10.5 percent of persons with unknown sex had the same mean lifespan as the remaining 89.5 percent of individuals with identified sex pooled together, so the existence of possible sex bias after sex identification looks unlikely. This data sample of 190,696 individuals with known sex out of 213,174 individuals was used for a more detailed study of gender-specific mortality at advanced ages.

The SSA DMF also contains data on month of birth for each person (in the overwhelming majority of cases), so it is possible to estimate life expectancy at age 80 years for each month of birth. In cohort life tables, which we have studied here, mean lifespan (mean age at death in a cohort) is equivalent to the life expectancy, while this is not the case for cross-sectional death records. We use the term life expectancy here instead of mean age at death in order to avoid confusion with mean age at death calculated on the basis of cross-sectional (death certificates) data. In order to avoid possible truncation biases, we estimated life expectancy in the age range of 80-110 years.

III. RESULTS

Quality of centenarian genealogical data

Our data consistency checks revealed a surprisingly small number of obvious data inconsistencies. In one case, an alleged centenarian had parents with incorrect birth dates (parents born later than the person himself). This case was dropped from the database. In another case, the centenarian's father was rather old (62 years) when the centenarian was born. This is not an impossible situation, so this case was left for further validation (this case was later confirmed through the SSA DMF but not found in early censuses and therefore not included in final analyses). All other records did not reveal obvious inconsistencies in event dates, so 990 records were left for further verification.

The overwhelming majority of genealogical records, when linked to the DMF, had revealed an identical birth and death year as well as birth and death month in both databases (687 out of 764 cases, or 89.9 percent). These matched records were additionally verified using information about first and last names (or last names of spouses for women), places of death (in the genealogy database) and places of last residence (in the DMF). When months of birth or death and years of birth or death did not match, then potential matches were established using information about place of death (in the genealogical file) and place of last residence (in the DMF). Thus, in addition to 687 (out of 990) persons having exactly the same birth and death dates in both databases, it was possible to add some records with birth or death dates that were not exactly identical in the genealogies and DMF. In most cases these differences were related only to minor disagreements in *exact month of birth or death* (not year), and in 731 cases (96 percent), the death year was the same both in the genealogy records and the DMF. One problem encountered for successful linkage to the DMF was surname changes by women after marriage. We resolved this problem by using surnames of spouses, which are available in the genealogical database, so that linkage success was approximately the same in both sexes. The number of successful links strongly depends on a centenarian's year of birth; persons born before 1890 were less likely to be found in the DMF (see Table 1 and Table 2 below). This result is consistent with previous reports that quality and coverage of the DMF database was lower for persons born before 1890 (Faig, 2001).

Table 1. Data on centenarians born in 1875-1900
Linkage success rate with the Social Security Death Master File (DMF)

Sex	Found in DMF	Total number of persons	Percent found
M	207	275	75%
F	560	715	78%
Total	767	990	77%

Thus, the proportion of successful links is 75 percent for males and 78 percent for females in total. For centenarians born after 1889, the percentage of successful links is higher at 82 percent. Among the 767 persons found in the DMF, centenarian status was confirmed in 744 of the cases.

734 centenarians had the same calendar year of *death* both in genealogy records and in the DMF. 714 centenarians had *both the birth and death years* identical in genealogy records and the DMF. Centenarian status could not be confirmed for only 23 alleged centenarians from the computerized genealogies (3.0 percent). The detailed breakdown of records found in the DMF is presented in Tables 3 and 4.

Table 2. Data on centenarians born in 1890-1900
Linkage success rate with the Social Security Death Master File (DMF)

Sex	Found in DMF	Total number of persons	Percent found
M	130	160	81%
F	421	511	82%
Total	551	671	82%

Table 3. Results of centenarian death dates verification using the Social Security Administration Death Master File (DMF)

	All centenarians born in 1875-1900	Centenarians born in 1890-1900
Total found in DMF	767	551
Centenarian status confirmed	744	535
Death year is exactly the same in genealogy and DMF	734	527

Note that the overall linkage success rate to the SSA DMF was moderate at 75-78 percent (Table 1). Also note (Table 4) that in 17 cases (2 percent) the difference between the death year in the genealogy records and the DMF was expressed in round numbers (e.g., 10, 20 or 30 years), which seems to be caused by misprints in genealogies. Thus all cases of exceptional longevity in genealogies should be verified using the SSA data.

Table 4. Comparison of death year reporting in genealogy and the Social Security Death Master File (DMF)

Age at death reported in genealogy	Number of cases	Difference between death year reported in genealogy and DMF
100	1	-1
	216	0
	1	1
	6	10
	1	20
101	4	-1
	241	0
	4	1

Table 4. Continued

Age at death reported in genealogy	Number of cases	Difference between death year reported in genealogy and DMF
101	2	10
	3	20
102	2	-1
	154	0
	1	2
	1	20
103	73	0
	1	1
	1	22
104	23	0
105	9	0
106	12	0
	2	20
107	5	0
	1	17
109	1	0
110	1	30
114	1	30

The lack of a match with the DMF could occur for a number of reasons: a misprint in genealogy, missing Social Security record (particularly if the person did not use Medicare benefits), difficulty in matching a person with a common name when the dates are not identical, etc. In addition, DMF covers about 90 percent of all deaths for which death certificates are issued (see Faig, 2001) and about 92-96 percent of deaths for persons older than 65 years (Hill, Rosenwaike, 2001). Further work with non-matched cases using additional data sources (obituaries, state collections of death certificates) revealed that about half of the non-matched cases are related to misprints in genealogies, and about 20 percent of the non-matched cases have correct death dates (as found by linkage to state death indexes) but are not recorded in the SSA DMF.

It should be noted that the linkage success rate to the DMF was substantially higher for persons born after 1889—at 82 percent. The 534 records for persons with confirmed centenarian status born after 1889 and matched to the DMF were used further in verification of centenarian birth dates through linkage to early censuses.

Verification of centenarian birth dates using the early U.S. censuses

The overall success rate for linkage of centenarian records to early U.S. censuses was 91 percent. Table 5 shows the results of record linkage to the early U.S. censuses.

The agreement between years of birth recorded in computerized genealogies and years of birth reported by the 1900 census as well as age reported by the 1910 census was surprisingly good; there was complete agreement in birth year between genealogy records and census records in 92 percent of cases. In one case only, the centenarian's year of birth was three years less than in the genealogy record, i.e., the centenarian was in fact *older* than was reported in the genealogy file. In 4.5 percent of cases, the birth year of the centenarian in the U.S. census was one year less than the birth year indicated in the genealogy database and in 3.5 percent of cases the centenarian was one year younger than reported in genealogy records. Disagreements between birth years reported in the censuses and genealogies were more notable for parents (about 15 percent of all cases) than for children, but in the majority of cases the differences did not exceed one year.

Table 5. Number and percentage of genealogical records that were successfully linked to early U.S. census records among records confirmed through linking to the Social Security Administration Death Master File (DMF)

U.S. census	Males		Females		Both sexes	
	Number linked to early census record	Percentage linked to early census record	Number linked to early census record	Percentage linked to early census record	Number linked to early census record	Percentage linked to early census record
1900	90	78%	292	79%	382	79%
1910	24	21%	76	20%	100	20%
1920	1	1%	2	1%	3	1%
Total	**115**	**100%**	**370**	**100%**	**485**	**100%**

As a result of this record linkage study, we could verify birth dates for 485 centenarians born after 1889. The steps of age verification for this group of centenarians are presented in Table 6.

Table 6. Summary of results of genealogy records' linkage first to the Social Security Administration Death Master File (DMF) and then to the early U.S. censuses

Steps of data verification	Number of records for centenarians born after 1889		
	Males	Females	Both sexes
Initial number of records	160 (100%)	511 (100%)	671 (100%)
Found in the DMF	130 (81%)	421 (82%)	551 (82%)
Found in the early censuses	115 (72%)	370 (72%)	485 (72%)

Thus, we obtained 485 records for centenarians with verified birth dates, confirmed centenarian status and detailed genealogies. We did not find many cases of significant age exaggeration among centenarians with known genealogies and verified death dates. In other words, the birth year is recorded more accurately in genealogies than the death year. The 25 cases of one-year discrepancy with the census records are more likely caused by inaccurate birth date reporting during census enumeration rather than inaccuracy of genealogical records. Most genealogical records provide a detailed date of birth (day, month and year) taken from birth certificates or family bible records while census records are based on verbal reports during enumeration.

Description of the validated data sample of centenarians

As a result of this validation study, a cohort of 485 centenarians born in the United States in 1890-1900 was identified. A general overview of the data collection, verification and linkage used for identification of these 485 cases is presented in Figure 1.

All centenarians had verified dates of birth and death and known information for parents, siblings, spouses and other relatives. Table 7 below shows the age and sex breakdown of centenarians with verified ages. This developed database on long-lived persons combines information on family characteristics with data on the early-life conditions taken from the 1900-1910 U.S. censuses. This database was used to test a number of hypotheses on the factors affecting exceptional longevity (presented later).

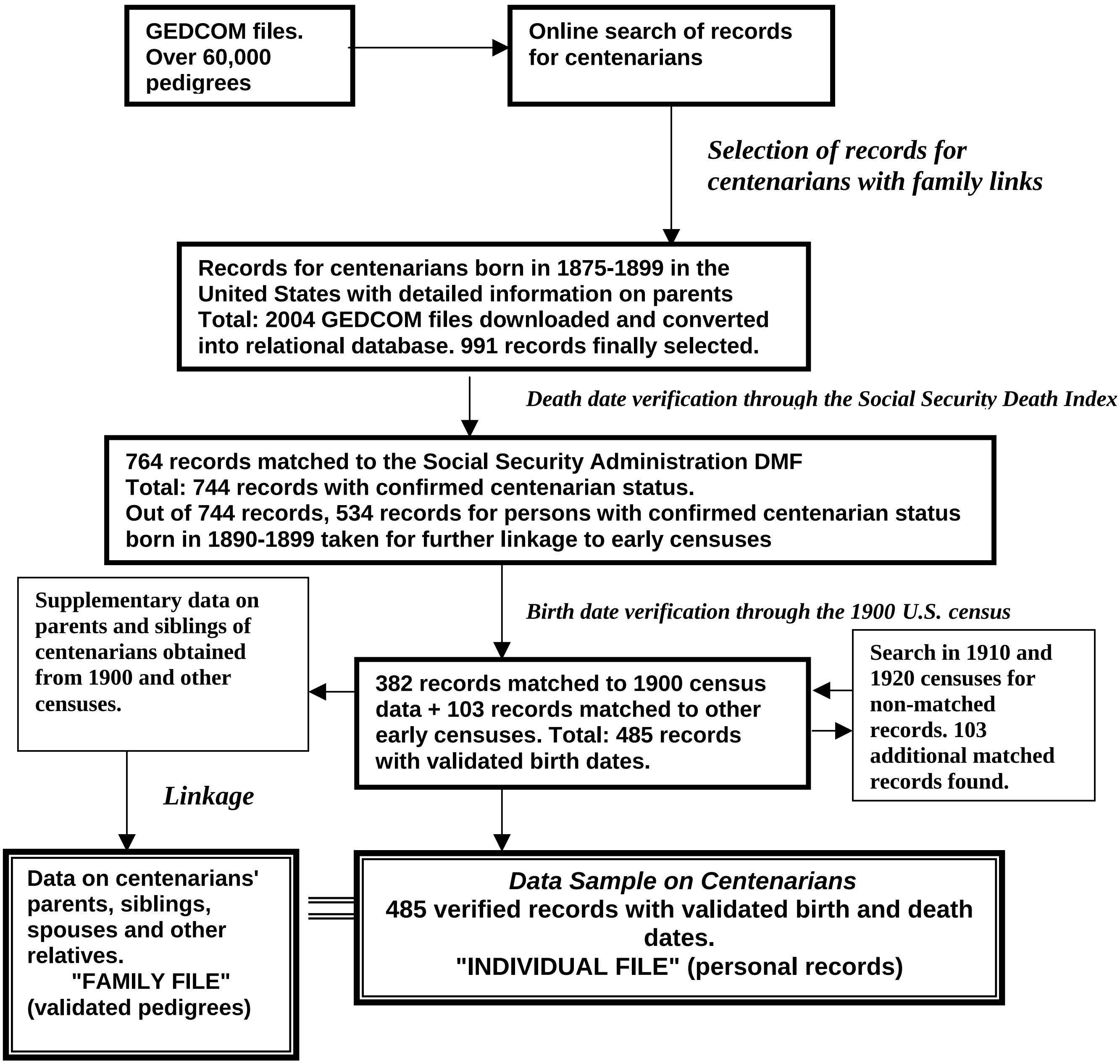

Figure 1. General overview of data collection and data processing protocol.

Beginning in the upper left, we searched the genealogical database Ancestry.com. Then, records for centenarian individuals born in 1875-99 in the United States with detailed information on both parents and grandparents were selected for further verification and analysis.

Table 7. Distribution of centenarians confirmed through linkage to the Social Security Administration Death Master File and early U.S. censuses, by age and sex.

Age at death reported in genealogy	Males	Females
100	40	102
101	34	134
102	28	83
103	9	37
104	2	7
105	1	0
106	1	6
107	0	1
Total	115	370

Birth order and survival to age 100: A study of within-family effects

The combined genealogical and census-related database was used in the analyses of several factors, which might be potentially important for survival to advanced ages. In particular, we focused our attention on the role of early-life living conditions in determining later-life survival.

Our previous studies on European aristocratic families lead us to a prediction that there might be a sex-specific link between a person's birth order and exceptional longevity. Specifically, women are more likely to become centenarians if they are born earlier than their siblings, when their parents are relatively young. This birth-order effect is expected to be stronger in women as compared to men. These predictions follow our earlier published findings that daughters conceived to older fathers live shorter lives, while sons are not affected by late conception (Gavrilov, Gavrilova, 2000; Gavrilov et al., 2003b; Gavrilova et al., 2003). Thus, if this prediction is correct, then the ratio of [*centenarian birth order/(family size + 1)*] should be significantly below 0.5 in females, but not in males. Note that testing this hypothesis is based on within-family analysis, which allows us to relax concerns over possible confounding effects of many other predictor variables that are fixed within each family (like parental lifespan, etc.).

According to our database, centenarians are usually born in rather large families with the mean number of children equal to 7.17 ± 0.17 (data on 392 families). For comparison, the mean number of children in the total sample of 29,118 families taken from the same genealogical sources and for the same historical period is 5.65 ± 0.02 children. Thus, centenarians tend to be born in larger families on average. Further studies are required to find out whether this is an important, meaningful finding or a trivial observation caused by ascertainment bias (according to probability theory, the chances that at least one child becomes a centenarian are higher for larger families with many children, simply because each child represents an additional "trial" for beating a longevity record).

To study the birth order effects, we have to remove non-informative cases where family size is equal to one and cases with less reliable information on family size (there are a few genealogies where family size was lower than that reported in the census).

The results of data analyses are presented in Table 8 below.

Note that the centenarian birth order ratio for female centenarians is indeed lower (0.45 ± 0.01) than expected (0.5) and this effect is statistically significant (P<0.01). In other words, the birth order ratio of centenarian women is 12 percent lower on average than it would be expected by pure chance (random uniform distribution for cases of exceptional longevity by birth order).

Table 8. Mean centenarian birth order *ratio* for male and female centenarians

Gender of centenarian	Number of cases (studied families)	Mean value of birth order ratio	Standard error	95% confidence interval	
Males	114	0.479	0.022	0.435	0.523
Females	370	0.446	0.012	0.423	0.470

Thus, female centenarians can be found less likely among later-born siblings conceived to relatively old parents. In contrast to females, the birth order ratio for centenarian men (0.48 ± 0.02) is closer to the theoretically predicted value of 0.5, suggesting that birth order is less important for exceptional male longevity (however, see the more detailed analysis provided later).

Similar results are obtained using another statistic named 'centenarian birth order difference.' If centenarians are distributed randomly by birth order within a sibship (independently of their centenarian status), then this difference should be equal to zero on average. This is what we expect to find for centenarian males, while this birth order difference should be negative for centenarian women, if the tested hypothesis is correct. The results of the data analyses are presented in Table 9.

Table 9. Mean centenarian birth order *difference* for male and female centenarians

Gender of centenarian	Number of cases (studied families)	Mean value of birth order difference	Standard error	95% confidence interval	
Males	114	-0.13	0.22	-0.55	0.30
Females	370	-0.50	0.11	-0.72	-0.28

Note that the mean value of the centenarian birth order difference for females is lower (-0.50 ± 0.11) than zero, and this difference is statistically significant ($P<0.01$). In other words, the birth order of centenarian women is lower on average than it would be expected (if birth order is irrelevant for longevity), and this difference in absolute terms corresponds to the shift of 0.5 to lower birth order on average. Thus, there is a tendency for female centenarians to be born among the first half of siblings in the family. In contrast to centenarian women, the birth order difference for centenarian males (-0.13 ± 0.22) is close to the theoretically predicted zero value, suggesting that birth order is less important for exceptional male longevity (however, we will present more detailed analysis later).

The results presented above are based on summary statistics, which describe the overall shift in birth order ranking of centenarians relative to their siblings. Deeper analysis of the odds to become a centenarian as a function of birth order (with centenarian siblings born in the same time window used as a control group) found that the best fit of the data both for males and females analyzed separately could be achieved with the following model:

$$\textbf{Logit (Longevity odds ratio)} = \boldsymbol{a\,x + b\,x^2 + c\,z + d},$$

where x is the birth order, z is family size, and a, b, c, and d are the parameters of the polynomial regression model. The choice of a quadratic model for birth order effect was made after studying all possible interactions between predictor variables. This study found that the quadratic effects of birth order are statistically significant, and therefore they cannot be ignored. On the other hand, the effects of higher order (cubic function) proved to be statistically insignificant, and they were dropped from the final model. Other interaction terms between predictor variables (birth order and family size) were found to be statistically insignificant and therefore were not included in the model.

The effect of family size, parameter *c*, was negative both for males (-0.11 ± 0.05, p = 0.028) and females (-0.07 ± 0.02, p = 0.002), which indicates that the odds of longevity are in fact decreasing in larger families. Further studies are required to find out whether this is a meaningful finding or a trivial consequence of ascertainment bias (the proportion of centenarians in a family is bound to decrease with increasing family size, because other siblings are likely not to be centenarians).

Figure 2 presents the results of data analysis in graphic form. It shows the dependence of the odds of living to age 100 as a function of a person's birth order (as predicted by the fitted polynomial logistic model). The graphs are computed for a fixed family size of 10 children (which is not particularly important, because family size influences only the vertical location of the curves rather than their shape because there is no interaction of family size with birth order).

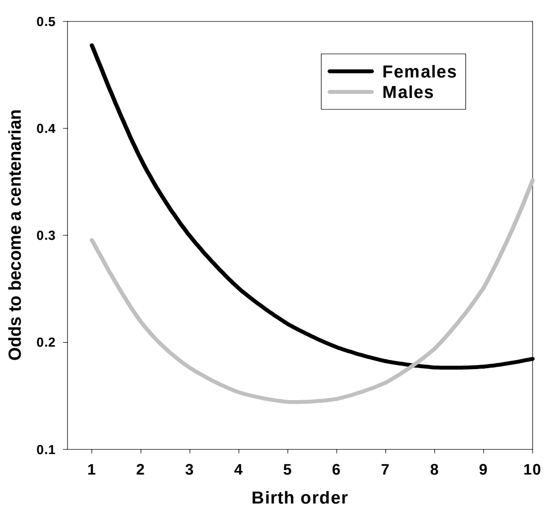

Figure 2.
The dependence of odds to become a centenarian on person's birth order as predicted by the fitted polynomial logistic model.

Note that the odds of becoming a centenarian decrease with birth order for females, which is consistent with the results of earlier data analysis based on summary measures. First-born daughters are almost three times (2.7 times) more likely to survive to age 100 compared to later-born daughters of higher birth orders (8+). Note that the strongest effect of birth order is observed when it is relatively small—one to five (see Figure 2), and then the birth order effect fades out.

The picture is different for males; there is an unusual U-shaped curve for the odds of living to age 100 in relation to birth order. The chances for exceptional longevity are minimal for sons having a birth order of four to six compared to those born earlier or later. Thus the earlier studies based on summary measures, which found no birth order effect in males, seemed to overlook it, because of a complex U-shaped form of the birth order effects in males. In fact, first-born sons are twice (2.05) as likely to become centenarians compared to sons having a birth order of five (see Figure 2). However, last-born sons (birth order 10+) also have more than twice (2.4 times) higher chances of surviving to 100 years compared to sons having a birth order of five. It is needless to say that these preliminary findings need to be replicated with other methods and datasets. For more detailed discussion of our findings see the "Discussion" section of this report (pages 55-59).

Early-life conditions and survival to age 100

The resulting dataset of 1900 and 1910 census households linked to the centenarian genealogies allows us to make a comparison of these households to the general set of households enumerated in early censuses available as part of the IPUMS from the University of Minnesota (IPUMS, see the Materials and Methods section). We conducted a comparison of households that raised a child-future centenarian (linked to the 1900 census) to the general sample of Caucasian households enumerated by the 1900 census, which had children below age 10 (to make these households comparable to our set of centenarians who were born in 1890-1899 and hence were below age 10 in 1900).

Our tested hypothesis is that if early childhood conditions are important for survival to age 100, then the households with children-future centenarians would be different from the general population. Tables 10 and 11 present results from multivariate logistic regression that estimate the odds for the household to be in the "centenarian" group. We conducted our analyses separately for male and female centenarians because our previous analyses demonstrated that men and women may respond differently to early-life living conditions (Gavrilov, Gavrilova, 1999; 2003a).

Data presented in Tables 10 and 11 demonstrate that both the region of residence and the household property status are the two most significant variables that affect the chances of a household producing a future centenarian (for both sons and daughters). We would assume that these chances are related to the chances of survival to age 100. Thus, spending a childhood in the Mountain Pacific and West Pacific regions in the U.S. may highly increase chances of long life (by a factor of 3) compared to the Northeastern part of the country. Also farm (particularly owned farm) residence results in better survival to advanced ages. This result is consistent with studies of childhood conditions and survival to age 85+ (Preston et al., 1998; Hill et al., 2000). These earlier studies, also based on linkage to early censuses, demonstrated a significant advantage in survival for children living on farms for both African Americans (Preston et al., 1998) and native-born Caucasians (Hill et al., 2000). On the other hand, the Northeast and Midwest were found to be the best regions for survival to age 85+ (Hill et al., 2000).

Table 10. Odds for household to be in the "centenarian" group for selected characteristics in the 1900 U.S. census. Female centenarians.

Characteristic	Odds ratio	p-value	95% confidence interval	
Census region:				
New England and Middle Atlantic	1.00 – reference level			
Mountain West and Pacific West	3.16	0.000	1.81	5.52
South (Southeast and Southwest)	2.05	0.002	1.30	3.23
North Central	2.42	0.000	1.58	3.70
Characteristics of father				
Immigration status				
Father immigrated	0.70	0.035	0.50	0.98
Father native-born	1.00 – reference level			
Literacy				
Father literate (can write)	1.29	0.352	0.76	2.19
Father illiterate	1.00 – reference level			
Survival of siblings in childhood:				
All mother's children survived	1.02	0.917	0.75	1.37
71-99% of children survived	1.00 – reference level			
Less than 70% of children survived	0.85	0.434	0.57	1.27
Household properties:				
Owned farm	1.00 – reference level			
Rented farm	0.63	0.007	0.45	0.88
Owned house	0.62	0.003	0.45	0.85
Rented house	0.26	0.000	0.18	0.37

Table 11. Odds for household to be in the "centenarian" group for selected characteristics in the U.S. 1900 census. Male centenarians.

Characteristic	Odds ratio	p-value	95% confidence interval	
Census region:				
New England and Middle Atlantic	1.00 – reference level			
Mountain and Pacific West	2.68	0.041	1.04	6.90
South (Southeast and Southwest)	1.11	0.797	0.51	2.41
North Central	1.39	0.372	0.67	2.89
Characteristics of father:				
Immigration status				
Father immigrated	0.40	0.019	0.19	0.86
Father native-born	1.00 – reference level			
Literacy				
Father literate (can write)	1.39	0.579	0.50	3.87
Father illiterate	1.00 – reference level			
Survival of siblings in childhood:				
All mother's children survived	0.91	0.734	0.51	1.54
71-99% of children survived	1.00 – reference level			
Less than 70% of children survived	0.93	0.848	0.45	1.91
Household properties:				
Owned farm	1.00 – reference level			
Rented farm	0.60	0.193	0.33	1.11
Owned house	0.28	0.001	0.13	0.58
Rented house	0.20	0.000	0.10	0.40

Both of the above-mentioned studies of childhood conditions and later survival found that a father's illiteracy significantly diminishes the chances of survival to age 85+. We find no such relationship for survival to age 100 in our dataset.

Having a father immigrant decreases the chances to become a centenarian for males and females, although for females the effect is weaker. A similar negative effect of the father's immigrant status was found for native-born Caucasians, both sexes combined (Hill et al., 2000). These findings do not support a hypothesis that healthier immigrants have healthier children, thereby explaining lower old-age mortality in the United States compared to other developed countries (Manton, Vaupel, 1995). Recently Costa and Lahey (2003) came to the same conclusion that immigration status is not related to better health. Finally, we found that deaths of siblings early in life had no statistically significant effect on the chances of becoming a centenarian. A previous study found that death of siblings decreases chances to survive to age 85 among African Americans (Preston et al., 1998). That study used more sophisticated methods of child mortality estimates (child mortality index) and copy-pair controls. In our study we used a proportion of surviving children reported by the mother during census enumeration as a proxy for child mortality within the household, and compared households where centenarians were raised, with a general population. For a more detailed discussion of our findings see the "Discussion" section of this report (pp.59-62).

Early-life conditions and survival beyond age 100

Our database on centenarians has individuals with verified lifespans from 100 to 107 years (see Table 7). Several claims of extraordinary longevity (over 107 years) were initially found in genealogies but they were not verified using the SSA DMF and the early U.S. censuses. The data sample for centenarians contains a number of variables describing early childhood conditions taken from early U.S. censuses. Other variables for regression analysis were taken from genealogical records: centenarian's month of birth, paternal and maternal lifespan, paternal and maternal ages at person's birth, family size (number of siblings) and birth year. We found that all variables describing childhood conditions in early censuses have no effect on survival beyond age 100 (data not shown). Although months of birth had some marginal effects on survival after age 100, the overall model was not statistically significant. These results demonstrate that mortality at such extreme ages is less dependent on the past and appears to be more sensitive to the current conditions. Further studies of extinct birth cohorts for centenarians born before 1890 could shed more light on this problem in the future.

Comparison of familial and sporadic centenarians

In this exploratory study we also tested a hypothesis that those centenarians who have family history of longevity (both parents lived to over age 80) are different from "sporadic" centenarians whose parents both lived to less than 80. This hypothesis is based on the idea that persons having long-lived parents may be less vulnerable to adverse environmental conditions (including early-life conditions) because of a favorable genetic background. As a result of these inherited longevity genes, offspring of long-lived individuals may survive to advanced ages despite poor childhood conditions because of additional protection provided by these genes . Some researchers believe that longevity is determined entirely by genetic background, while environmental factors like diet, educational background and economic status do not play a significant role in the ability to survive to advanced ages (see Perls, 2002).

This hypothesis was tested using the logistic regression model with a binary outcome variable indicating whether a centenarian represents a sporadic or a familial case of longevity. This model allowed us to test the hypothesis of whether childhood conditions (farm background, living in the West, etc.) were more important for becoming a sporadic centenarian (without having a familial longevity background). We found that both groups of centenarians do not differ from each other regarding the childhood conditions reported in early U.S. censuses as well as month of birth, birth order and family size (data not shown). Thus, farm background and other early-life conditions explored in this study are of similar importance for both groups of centenarians, which indicates that early-life conditions are important for survival to advanced ages, even for persons with a favorable genetic background.

Mortality beyond age 100: Evidence from the SSA Death Master File (DMF)

The collection of records from the SSA DM F undertaken in this validation study had some interesting ramifications for mortality analyses at advanced ages.

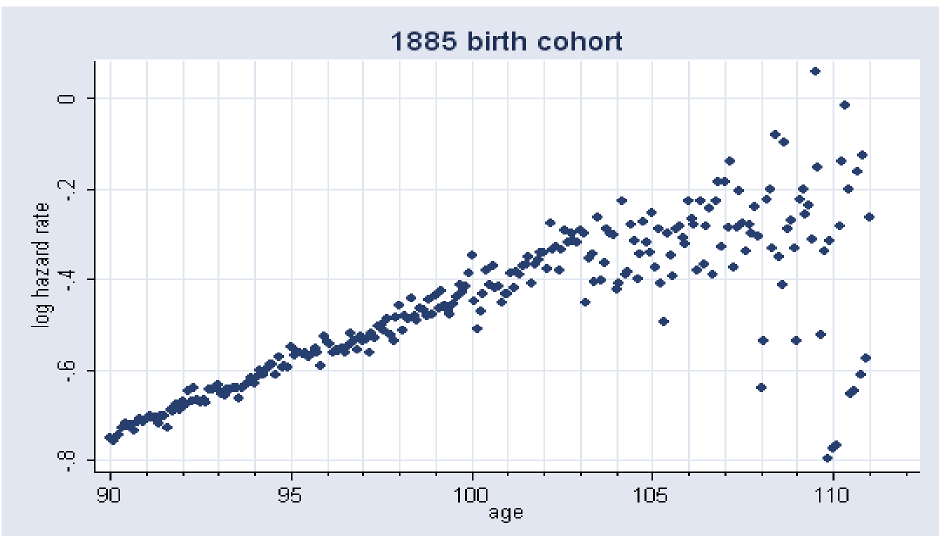

Figure 3. Hazard rate (mortality force, year^{-1}) for 1885 birth cohort.
Data from the Social Security Administration Death Master File. Total U.S. population.

First, we estimated hazard rates for single-year extinct birth cohorts at each month for ages over 90 years. Results of the hazard rate estimates for three birth cohorts (1885, 1889 and 1891) are presented in Figures 3-6.

A recent study of age validation among supercentenarians (Rosenwaike, Stone, 2003) showed that age reporting among supercentenarians in the SSA database is rather accurate, with the exception of persons born in the Southern states. In order to improve the quality of our dataset when estimating hazard rates, we excluded records for those persons who applied for Social Security numbers in the Southeast (AR, AL, GA, MS, LA, TN, FL, KY, SC, NC, VA, WV) and Southwest (AZ, NM, TX, OK) regions, Puerto Rico and Hawaii.

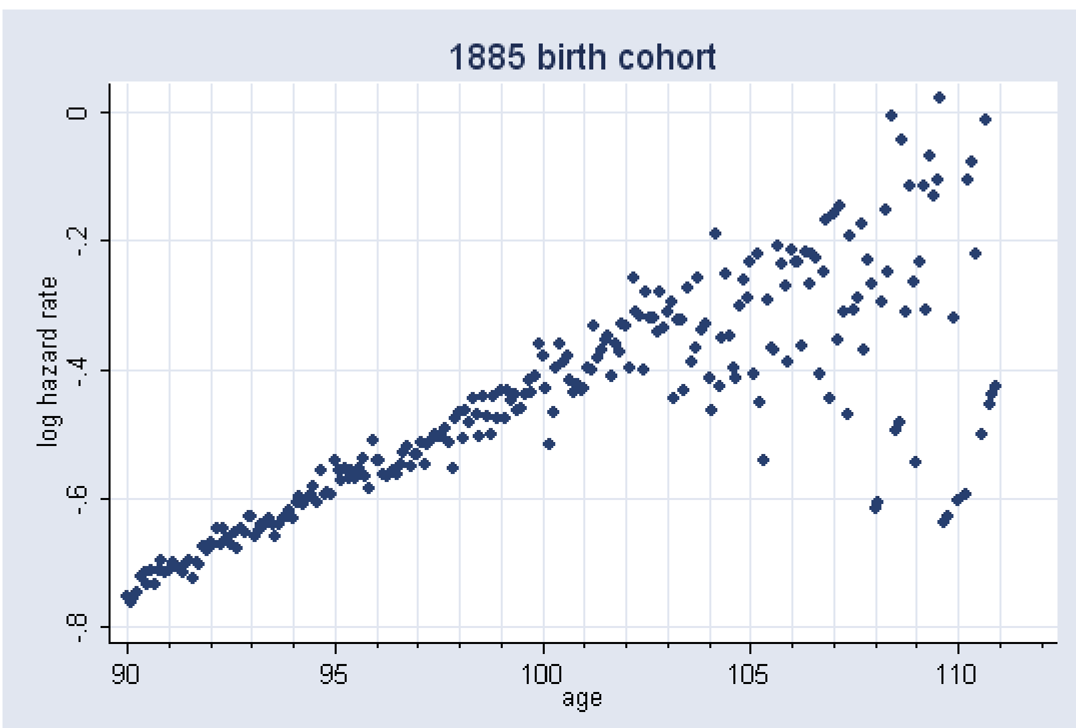

Figure 4. Hazard rate (mortality force, year^{-1}) for 1885 birth cohort.
Data from the Social Security Administration Death Master File.
Less reliable data for Southern states, Puerto Rico and Hawaii are excluded.

This step of data cleaning however, did not significantly change the overall trajectory of mortality at advanced ages, but decreased the number of too-low mortality estimates and increased the number of higher-mortality estimates after age 105 (see Figures 3-4).

Note that up to ages 102-105 years, mortality grows steadily without obvious deceleration. Only after age 105 does mortality tend to decelerate, although high statistical noise makes mortality estimates beyond age 105 less reliable (also note that for cohorts born after 1890, mortality over age 110 is affected by data truncation). These figures demonstrate that single-year birth cohort mortality agrees well with the Gompertz law up to very advanced ages.

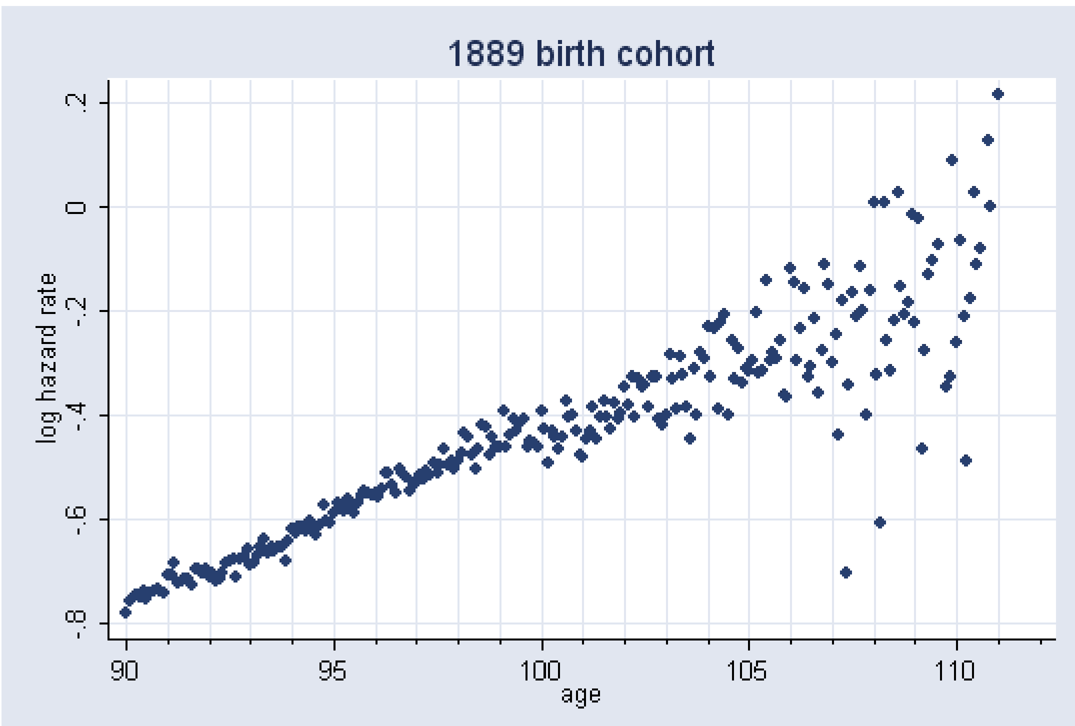

Figure 5. Hazard rate (mortality force, year^{-1}) for 1889 birth cohort.
Less reliable data for Southern states, Puerto Rico and Hawaii are excluded.

Previous studies of mortality at advanced ages used aggregated data, combining several birth cohorts with different mortality, and this aggregation of heterogeneous data could produce mortality deceleration and subsequent leveling-off, as it is predicted by the heterogeneity model (Beard, 1971). Mortality deceleration and even decline of mortality often are observed for data with low quality. On the other hand, improvement of data quality results in straighter mortality trajectory in semi-log scale (Kestenbaum, Ferguson, 2001). In our study, the more recent 1891 birth cohort demonstrates straighter trajectory and lower statistical noise after age 105 than the older 1885 one (see Figures 4 and 6). Thus, we may expect that cohorts born after 1891 would demonstrate an even better fit by the Gompertz model than the older ones because of the improved quality of age reporting. Testing this hypothesis now is hampered by the problem of data truncation for non-extinct birth cohorts, so we have to wait until these cohorts become extinct.

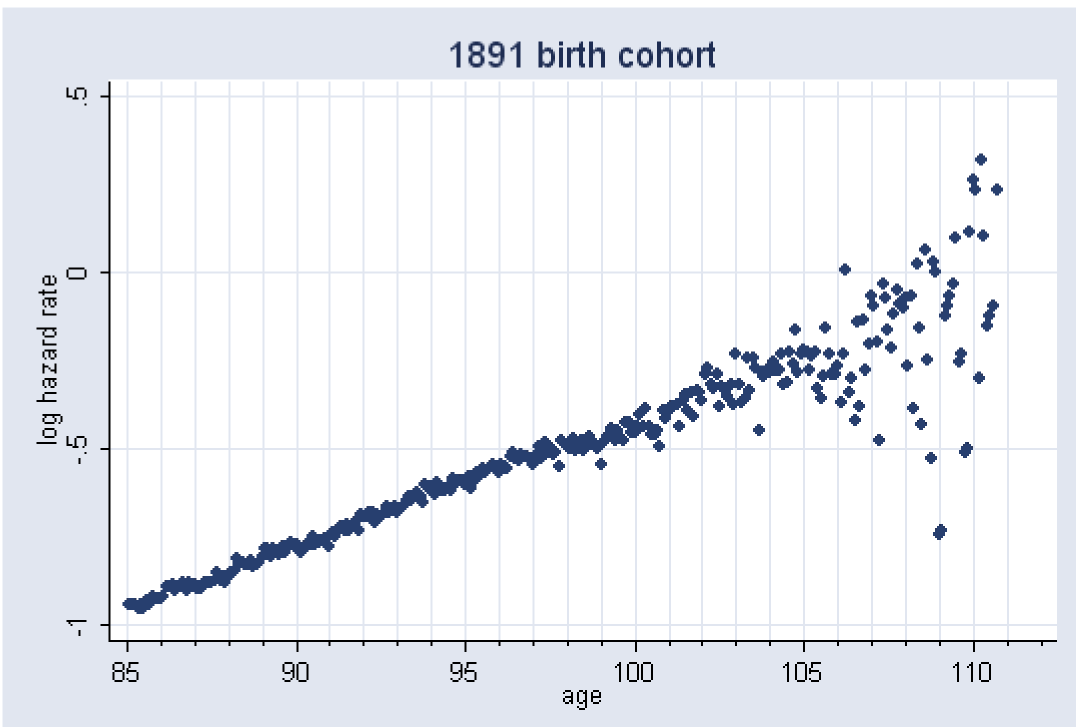

Figure 6. Hazard rate (mortality force, year^{-1}) for 1891 birth cohort.
Less reliable data for Southern states, Puerto Rico and Hawaii are excluded.
Data from the Social Security Administration Death Master File.

The result of hazard rate estimation for males and females is presented in Figure 7. Note that male mortality continues to exceed female mortality up to very advanced ages. At age 110 the number of remaining males (9 persons) and females (44 persons) is too small for accurate estimates of hazard rate after this age.

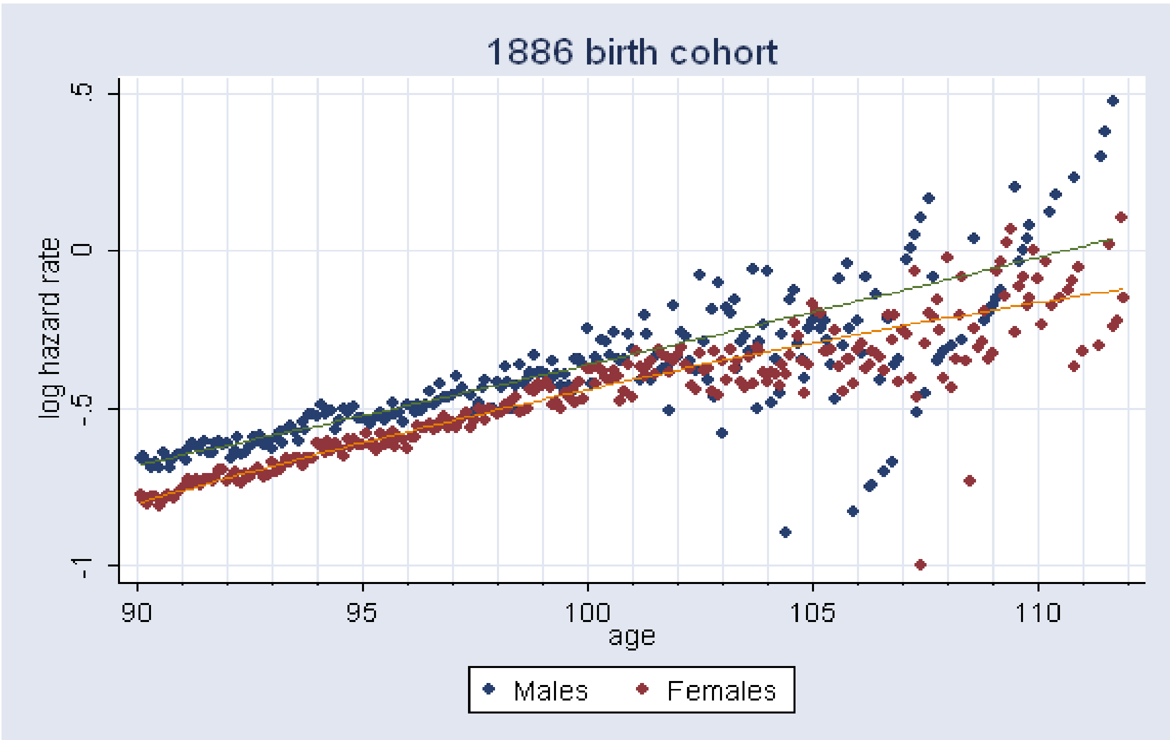

Figure 7.

Hazard rate (mortality force, year^{-1}) for males and females from 1886 birth cohort. Data are fitted using quadratic polynomial regression. Total U.S. population.

Interestingly, the hazard rate estimates made using crude estimates of lifespan in whole years (like in standard demographic life tables) create an appearance of more pronounced mortality deceleration (Figure 8) than estimates obtained for every month of age (Figure 7).

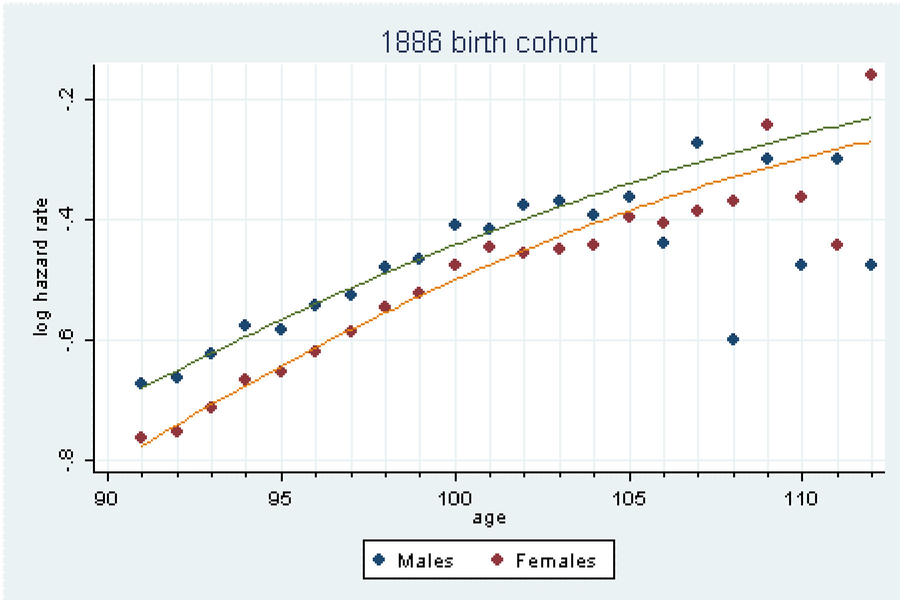

Figure 8.

Hazard rate (mortality force, year^{-1}) for males and females from 1886 birth cohort. Lifespan is estimated in whole years. Data are fitted using quadratic polynomial regression. Total U.S. population.

Overall these data demonstrate that mortality at advanced ages follows the Gompertz law up to the ages 102-105 years and these findings have an obvious actuarial significance.

The results obtained in this study are interesting, yet should be regarded with some caution. The SSA DMF provides no information about sex and race of decedents. Also, quality of data for earlier birth cohorts is lower than for more recent birth cohorts. Thus, we may expect that 5-10 years from now the quality of the SSA DMF data would be sufficient enough to obtain more accurate estimates of mortality at advanced ages.

Month of birth and mortality at advanced ages

The SSA DMF data allow us to explore another interesting problem: the effects of early-life conditions and month-of-birth in particular on mortality at advanced ages.

The SSA DMF contains data on month of birth for each person (in the overwhelming majority of cases), so it is possible to estimate life expectancy at age 80 for each month of birth. In cohort life tables, which we have studied here, mean lifespan (mean age at death in a cohort) is equivalent to the life expectancy, while this is not the case for cross-sectional death records.

Figure 9 shows the effects of month of birth on life expectancy at age 80 for two birth cohorts: 1885 and 1891. Note that persons born in January have higher life expectancy at age 80 than persons born in April-June (about 2 percent difference, statistically significant).

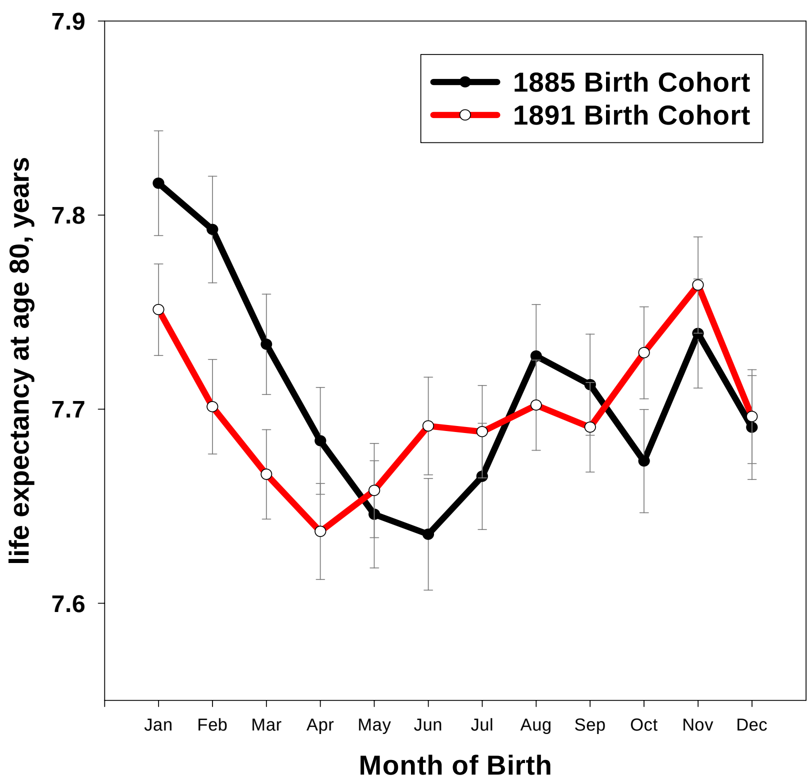

Figure 9. The dependence of life expectancy at age 80 on person's month of birth. Comparison of 1885 and 1891 birth cohorts. Data on extinct birth cohorts obtained from the SSA DMF. Total U.S. population.

Figure 10 confirms replicability of this observation for a longer time period: practically all single-year birth cohorts born from 1885 to 1899 demonstrate the same monthly pattern in life expectancy. It is interesting that monthly pattern does not change for this relatively long 14-year time period. Thus, life expectancy at age 80 depends on month of birth: persons born in January (or December) live longer than persons born in other months and in April-June in particular. This seasonal pattern repeats in every birth cohort from 1885 to 1899.

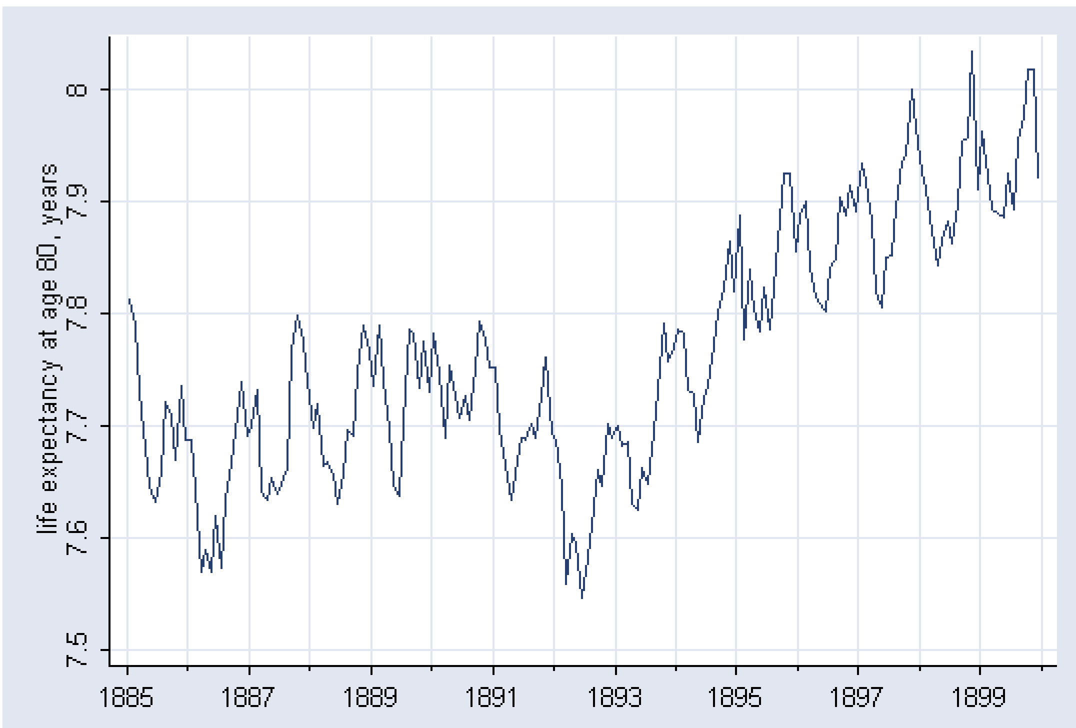

Figure 10. Periodic seasonal changes in life expectancy at age 80 for 1885-1899 birth cohorts depending on month of birth. Total U.S. population.

However, by age 100 this monthly pattern in life expectancy apparently disappears, indicating that centenarians indeed represent a selected population (Figures 11-12).

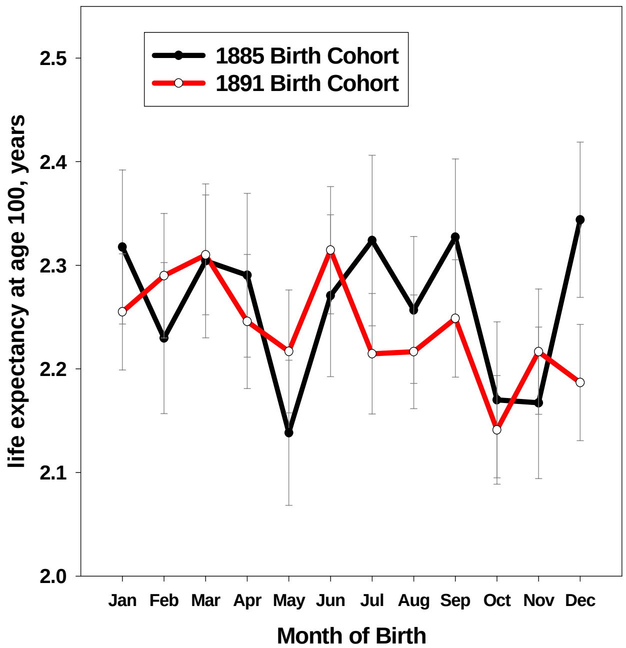

Figure 11.
Lack of seasonal dependence of life expectancy at age 100 on person's month of birth. Comparison of 1885 and 1891 birth cohorts. Total U.S. population.

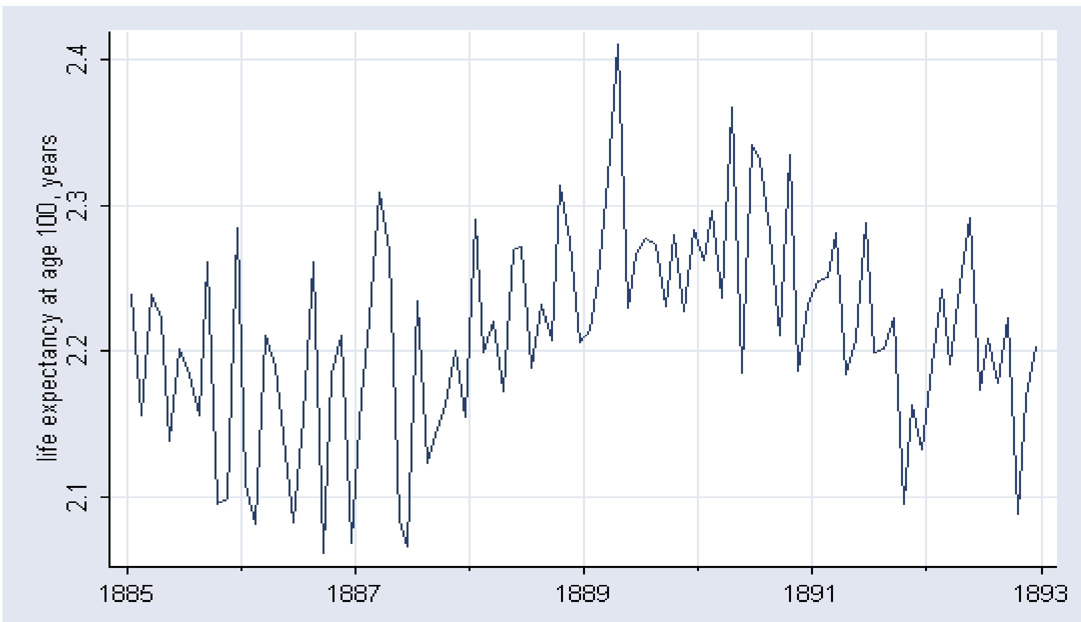

Figure 12. Life expectancy at age 100 for 1885-1893 birth cohorts. Total U.S. population.

We already demonstrated in our previous studies that the month-of-birth pattern of survival depends on age, so overall monthly patterns might be different in different periods of life. For example, in the study of 1800-1880 birth cohorts of European aristocracy, we found that lifespan at age 30 is particularly low for February-born women and higher for December-born ones (Gavrilov, Gavrilova, 1999). However this monthly pattern changed when lifespan at age 50 and over was analyzed (Gavrilov, Gavrilova, 2002).

IV. DISCUSSION

Computerized genealogies contain important information about family and life-course events, which are otherwise difficult to collect: lifespan of parents and other relatives, number and sex of siblings, birth order, ages of parents when person was born, age at marriage, number of spouses and lifespan of spouses and other non-blood relatives, number and sex of children and timing of their birth, place of birth and information about residence during the life-course (derived using places of birth for siblings and children). Thus, computerized genealogies may be a valuable resource for studies of mortality and longevity. However, the reliability and quality of computerized genealogies are uncertain, resulting in underutilization of this data resource by researchers. In this project we developed a technique of genealogical data collection, verification and utilization in scientific analyses of longevity.

This exploratory project demonstrated that the ongoing revolution in information technologies created unique opportunities for conducting actuarial studies. In particular, the online availability of early censuses greatly accelerated and facilitated the process of record linkage used in the process of centenarian age verification. In our study the overall matching success rate of linkage to early U.S. censuses was 91 percent, which is significantly higher than in other studies on linkage to early censuses: 39-56 percent (Rosenwaike, Logue, 1983; Guest, 1987; Rosenwaike et al., 1998), 69 percent (Hill et al., 2000) and 54 percent overall and 69 percent for Caucasians (Rosenwaike, Stone, 2003).

The reasons for the relatively high success rate of linkage to the early censuses in our study can be explained by the availability of detailed supplemental information in genealogical records. The most important piece of information for successful searches in census records was information on *places of birth* for siblings born close to the census date. Thus, if the family moved to another state after the birth of the alleged centenarian, his/her family could be easily traced using information about the birthplaces of other siblings. This is an important advantage compared to the traditional studies of record linkage to the early U.S. censuses based on information taken from the Social Security SS-5 forms (Rosenwaike et al., 1998; Hill et al., 2000; Rosenwaike, Stone, 2003).

We had no need to apply the scoring system of match rating suggested in previous studies (Hill et al., 2000; Rosenwaike, Stone, 2003), because the availability of supplemental information in genealogy records made the judgment about match or non-match perfectly clear. If the names and years of birth for parents and siblings are in a good agreement in both the genealogy database and census, the match is considered to be very confident. On the other hand, if the names of parents are the same in the census and genealogy, but siblings have different names, it is quite clear that the match is not acceptable. In some rare cases of small families with one or two children, additional information about places of birth for parents and children was used to resolve the problem. Unlike previous studies of linkage to early censuses, we did not encounter problems with persons having common first and last names because detailed information about place of birth for the potential centenarian and his/her siblings (state, county, township) helped to identify the correct match among many potential matches. The detailed information about names, ages and places of birth for parents and siblings available in genealogies helped us to avoid ambiguous matches, which should be common in linkage studies based only on the information about parental names and places of birth and residence (Rosenwaike et al., 1998). The main difficulty we encountered in our search was related to rare and unusual first names, which were spelled in a variety of ways in census indexes.

During the centenarian birth date verification process, we also tested a suggestion that deceased elder siblings of the same name might be incorrectly cited as centenarians in genealogies. Such cases of "identity theft" are well known in centenarian studies. For example, Pierre Joubert, who appeared in the Guinness Book as a 113-year-old man, in reality died at 65 years, whereas his namesake—his son—died 48 years later (see Jeune, Vaupel, 1999). Such a scenario, however, is highly unlikely when detailed genealogies are available, and it was a Canadian genealogist and demographer, Hubert Charbonneau, who demystified the Pierre Joubert case using genealogical methods. In complete and detailed genealogies, this scenario of "identity theft" looks highly unlikely. Almost all genealogies with families having deceased children (88 percent) reported all children, including those who died in infancy. Only in two out of 198 such families the younger child was named after his or her elder sibling (and this younger sibling was not a centenarian in both cases). Thus, the appearance of a centenarian with a false identity in the genealogy should involve a combination of three relatively rare events: naming a child after a deceased elder sibling, non-reporting the deceased child in the genealogy and survival of a sibling to an advanced age (even a younger sibling should become at least an octogenarian or nonagenarian). Thus, it seems that "identity theft" by centenarians is not a likely phenomenon in detailed and complete genealogies.

This study demonstrated that quality of computerized genealogies is good enough for conducting scientific research, if only the detailed and complete genealogies are selected. If birth dates and death dates of persons, as well as their parents, are available in the genealogy, then such genealogies might be considered to be a basis for further studies. We found that the quality of birth dates reporting in genealogies is particularly high. Frequency of serious misprints in death dates is higher, although even in this case it is close to 2 percent only. An internal consistency check is a good way to eliminate potential misprints in genealogies, and all cases of extreme longevity require validation.

Study of birth order effects using such aggregated summary measures as "centenarian birth order ratio or difference" demonstrated that women seem to be more likely to become centenarians if they are born earlier compared to other siblings, when their parents are relatively young. In contrast to women, the birth order of centenarian men initially seemed to be no different than what would be expected by pure chance. These observations corresponded well with earlier published findings obtained on European aristocratic families that daughters conceived to older fathers live shorter lives, while sons are not affected by the fact of their late conception (Gavrilov, Gavrilova, 2000; Gavrilov et al., 2003b; Gavrilova et al., 2003). Obviously the later-born children having a higher birth order are also the children conceived to older fathers (on average). Therefore, one can speculate that the birth order effects observed in this study may be related to paternal-age effects, which were already documented in scientific literature (cited earlier), and were speculated to be attributed to accumulation of deleterious mutations in parental germ (sperm) cells.

However a more detailed subsequent study on the effects of birth order revealed that the birth-order effects in males were simply overlooked because they proved to be not monotonic, as initially assumed, but rather U-shaped (see Figure 2).

For females, the birth order effects are high for low birth orders and then the birth order effects fade out. In other words, it is good for female longevity to be born among the first children, while for the last-born children the exact birth order is less important. It is obvious that these kinds of studies may have significant implications for actuarial science and practice.

A large number of studies have demonstrated that children with high birth order tend to have a disadvantaged position during childhood years with regard to both health (Sears et al., 1957; Nixon & Pearn, 1978; Kaplan et al., 1992; Elliot, 1992) and educational achievement (Belmont & Marolla, 1973; Belmont et al., 1976; Breland 1973, 1974).

Even if birth order *per se* may not have a direct influence on survival, there are some childhood living conditions important to health and mortality that are related to birth order. An excellent review of the existing hypotheses and facts regarding birth order effects is provided by Modin (2002), which is briefly summarized here.

One hypothesis, called "resource theory," assumes that family resources (both material as well as human) may become diluted as the family grows larger (Blake, 1981, 1989). Thus, in contrast to most first- and earlier-borns, children of high birth order are born into conditions characterized by more limited access to parental attention and supervision (Hanushek, 1992). Such limited access to parental care may also result in less attention being paid to the health and safety of these children during their first years of life. It was demonstrated that later born children in large families have a higher risk of accidents during early childhood (Nixon & Pearn, 1978; Bijur et al., 1988). Specifically, children of higher birth orders are more likely to experience accidents at ages between 1 and 5 years (Bijur et al., 1988). Also, among children who had experienced a non-fatal drowning accident, more than half were found to be last-born children of large sibships. The causes of these findings may be lack of time for parental supervision leading to an increased risk of accidents for later-born children of large sibships (Nixon & Pearn, 1978).

Lack of parental attention may result in less efforts invested in the health of the child and disease prevention. Earlier-born children receive more health care (Horwitz et al., 1985) than those of later birth order (Horwitz et al., 1985; Celik & Hotchkiss, 2000; Fergusson et al., 1984) including immunization against diphtheria, smallpox and polio (Kaplan et al., 1992). This relative lack of medical care early in life may have long-lasting consequences on health and mortality later in life.

Other factors that should be taken into account while studying the mortality and health of persons born in the late 19th and early 20th centuries are poverty and a high rate of infections at that time. In the early 20th century, later-born children living in poor families were more affected by the shortage of economic resources and crowded housing during early childhood than their earlier-born siblings. Higher-order children who were born into an already crowded household were at higher risk of infectious diseases, which were very common in the late 19th and early 20th centuries (Burnett, 1991). The higher mortality of later-born infants was found not only for historical data but also for contemporary data on Swedish and Norwegian infants born in 1985-88 (Espehaug et al., 1994).

Another hypothesis providing a possible explanation of birth-order effects is a hypothesis of the "biological depletion" of the mother, explaining the often-observed higher mortality of children born in the end of the mother's reproductive life (Chidambaram et al., 1987; Majumder, 1988).

Interestingly, the recent Swedish study of birth-order effects on adult survival at ages 20-54 years (Modin, 2002) has produced results that are similar to the findings presented here. Specifically, this study also found a U-shaped dependence of survival chances as a function of birth order. This study of adult mortality revealed that "*a hump-shaped association appears to exist for both men and women, with first and very late borns having approximately the same mortality risk*" (see p. 1059, Modin, 2002), while individuals with intermediate birth orders (3-6) had the highest risk of death at adult age (Modin, 2002). Because of the inverse relationship between mortality and survival, this "hump-shaped association" for mortality corresponds to the U-shaped association for survival chances described in our study. No good theoretical explanation had been suggested so far for this puzzling observation, but a speculation was made that "*having a large number of [older] siblings may well be considered a resource in many respects. Older brothers and sisters may serve as role models for younger siblings, and they are often important sources of social support*" (see p. 1051-1052, Modin, 2002). Therefore, "*it is possible that, at adult and old age, having a large number of [older] siblings acts as a buffer against ill health and mortality by means of greater access to social support from the family of origin*" (see p. 1059, Modin, 2002). Later-born children are also heavier at birth, on average, which is considered to provide a survival advantage (Magnus et al., 1985; cited in Modin, 2002). Heavier newborn children are less prone to many diseases in adult life (Barker, 1998). Another interesting observation of this Swedish study, which corresponds well with our findings, is a stronger birth-order effect in women when compared to men (Modin, 2002). This Swedish study also found that the birth-order effects on human mortality and survival are qualitatively different in different age groups, which adds to the complexity of this problem (Modin, 2002).

We believe that further more detailed studies are necessary to explore the mechanisms of birth-order effects. The importance of this exploratory study is in (1) detection of significant birth-order effects on chances of survival to extreme old ages; (2) demonstration that these birth-order effects may have a complex U-shaped dependence, and, therefore, may be overlooked in previous studies assuming monotonic dependence.

Study of other early-life living conditions. Studies of exceptional longevity using genealogical data require a choice of an appropriate control group. One approach is to use a population-based control group. We applied this approach in studies of early-life conditions and survival to age 100 using a control group taken from IPUMS (see earlier). Our findings agree with previous reports on the effects of childhood conditions on survival to advanced ages (Preston et al., 1998; Hill et al., 2000). However, we should admit certain limitations of our study. Comparison with population samples assumes that differential survival is the only cause of differences between cases and controls. However, in our case computerized genealogies of good quality may not represent a random representative population sample. The absence of African American centenarians is one obvious bias of our sample, which can be explained by difficulties in genealogy compiling because of the paucity of historical information for African Americans, lower popularity of this genealogical activity among African Americans and less attention to date and age recording (see Hill et al., 1995), which selected out potential African American genealogies during our initial screening. For other studied variables, the possibility of bias is not so certain. The proportion of genealogies compiled for families originating from the New England and Middle Atlantic regions is by no means lower than for families originating from the Western region, because of the much longer documented history of New England and East Coast families. There is no reason to believe that household characteristics are different for families covered by genealogies and the general Caucasian population. The definite answer to this question could be obtained by comparing computerized genealogies for "normal" (non-centenarian) individuals with population characteristics drawn from the IPUMS database, a study that we hope will be conducted in the future.

Data from early censuses linked to computerized genealogies add additional important information about conditions during a person's childhood. In this study we compared data for centenarians with population-based controls. This approach allowed us to study the effect of early place of residence on the chances of survival to advanced ages. In addition, there might be other approaches for the choice of control group. One approach is to select a family of neighbors enumerated on the same or adjacent page of an early census, which has a child of the same age, and use this family as a control. Another approach is to take a control group from the genealogy (e.g., sisters- or brothers-in-law) and link these individuals to early censuses as well. In both cases we relax a problem with potential bias caused by selection of genealogies, but lose an opportunity to study the geographical effects of early residence (cases and controls have the same or almost the same place of residence).

In general, our results support the idea that early childhood conditions might be important for survival to advanced ages (Gavrilova et al., 2003; Costa, Lahey, 2003). Possible mechanisms of these early-life effects were discussed at the international symposium, "Living to 100 and Beyond" where this study was presented. In particular, Thomas Edwalds has suggested the following useful comments, now published in the "Living to 100" monograph: "*The first [comment] concerns the Gavrilov study and [parental] farm ownership being a significant factor* *[for a child to survive to 100]. Without the type of food processing that's currently available, living on a farm 100 years ago meant fresher food with more nutrient value. It very well might correlate to prenatal and perinatal nutrition to have that as one of your significant factors predicting the mortality at advanced ages.*" (published online, see page 5 at: http://library.soa.org/library-pdf/m-li05-1_tr4.pdf).

Indeed, we found that the chances of becoming a centenarian are inversely related to the level of the sickness burden in early life (measured through the level of child mortality) in compared groups of U.S. populations. This pattern was found by comparing the results of an earlier study by Preston and Haines (1991) on child mortality in the 1900 U.S. population with the results of our study on longevity chances in the same groups of the U.S. population. Specifically, families of farm owners had lower child mortality (see p.113 in Preston and Haines, 1991), and more long-lived children (our study, see Tables 10 and 11) than families renting a home.

Also, children born in the North Atlantic region of the United States had higher child mortality in the 1900s compared to the Western region (see p.112 in Preston and Haines, 1991). These findings correspond well with our observation that children born in the Western region of the United States in the same time period are more likely to become centenarians later when compared to children born in the North Atlantic region (see Tables 10 and 11).

It is also interesting to mention that the highest body weight among the World War I recruits was observed among those U.S. recruits who were born in the Western United States (see p.114 in Preston and Haines, 1991). In other words, the heaviest recruits came from the West, and there was a significant negative correlation (r = -0.65) between recruits' body weight and levels of child mortality in the regions where recruits were born (see p.114 in Preston and Haines, 1991). These observations suggest that there were indeed large regional differences in the sickness burden, which in disadvantaged regions, led to higher mortality of children, their impaired growth (reflected in lower weights of recruits), and, as we found in this study, lower chances of surviving to 100 years. Obviously, this historical correlation between child mortality and weight attained by young adults may not be applicable to modern populations of industrialized countries because of growing obesity problems.

The study of mortality at advanced ages for four U.S. birth cohorts (1885, 1886, 1889 and 1891) showed that mortality steadily increases with age without significant deceleration from 90 to 105 years. Then statistical noise rapidly increases and mortality tends to decelerate. We already noted that the period of mortality deceleration in mammals is very short compared to lower organisms. Our study shows that it appears to be relatively short in humans, too. This observation agrees well with the prediction of the reliability theory of aging, according to which more complex living systems/organisms with many vital subsystems (like mammals) may experience a very short or no period of mortality plateau at advanced ages in contrast to simpler organisms (Gavrilov, Gavrilova, 1991; 2001b; 2003a).

This study also demonstrated that life expectancy at age 80 depends on month of birth: persons born April-June live shorter lives than persons born October-November, and this periodicity repeats in every birth cohort. These seasonal patterns are consistent with earlier reports that persons born in the second quarter live shorter lives than persons born in the fourth quarter (Costa, Lahey, 2003). These monthly patterns also partially agree with a previous study based on aggregated death certificates, which found the peak of mean age at death in September/October and the trough in June/July (Doblhammer, 2003). Agreement with results obtained on the basis of cross-sectional data might indicate that the effects of month of birth are indeed rather stable over time. This stability is evident at least for the 1885-1899 birth cohorts

(Figure 10). The fact that such an early circumstance of human life as the month of birth may have a significant effect 80 years later on the chances of human survival is quite remarkable. It indicates that there may be critical periods early in human life particularly sensitive to seasonal variations in living conditions in the past (e.g., vitamin supply, seasonal exposure to infectious diseases, etc.). This hypothesis explains why the effects of month of birth on human lifespan may have some actuarial significance.

V. CONCLUSIONS

This exploratory study has a number of new and interesting implications for actuarial science. In general, this study has demonstrated that an ongoing revolution in information technology and computer science has created new opportunities for actuarial studies on human longevity. Millions of individual records on human lifespans are now computerized and are available online (Social Security Administration Death Master File, genealogical records, etc.). Moreover, detailed information for each member of the entire population of the United States has become available online in the form of images of the early U.S. censuses, including the most recent publicly available 1930 U.S. census.

This study has demonstrated the opportunities of using these rich information resources for developing a reliable database for actuarial studies on human longevity. In this exploratory study, we found that the best way to start human longevity database development is to first use the family-linked data available in computerized genealogies.

We found that contrary to the common belief in the poor quality of genealogical data, this information resource is highly valuable if we follow certain methodological guidelines uncovered in this study.

These methodological guidelines are:

1. To use only those genealogical records that contain complete, exact and detailed dates of birth and death, places of birth and information on parental names and lifespans.
2. To use this genealogical data as a starting point only, subject to subsequent external validation with the Social Security Administration Death Master File and the early U.S. censuses.

Perhaps most important, a particular procedure of data matching and cross-checking has been applied in practice, which produced a reliable dataset with several hundred family-linked records for individuals with exceptional longevity.

Now, when a working procedure of database development is in place, it could be applied on a larger scale to get many thousands of family-linked records of exceptional human longevity, with obvious implications for actuarial science and practice.

Other implications of this study are related to the identified putative predictors of human longevity. It came as a surprise to us that the geography of a birth place (or factors associated with it) within the United States seems to be an important determinant of human longevity. Our preliminary findings suggest that there may be a threefold difference in chances of survival to age 100, depending on the location of childhood residence. Two kinds of implications are important here. The methodological implication is that future studies should not be limited to a common practice of using a geographically matched control group for comparison purposes, because this study design overlooks the importance of geographic factors. A substantive implication is that the mechanisms of this early-life location effect on human longevity need to be studied and understood, and the alternative trivial explanations (like selection bias) need to be excluded in future studies. Another interesting observation of this study is a very strong effect of farm background on survival to advanced ages, particularly for men.

Obtaining data from the SSA Death Master File allowed us to conduct some additional analyses of mortality at advanced ages, which may be important for actuarial science. First, it was possible to obtain estimates of mortality at extreme ages for more homogeneous single-year birth cohorts. It was found that cohort mortality follows the Gompertz law up to the age of 105 years, and the period of mortality deceleration is not as long as it was believed on the basis of period life tables. After age 105, the mortality estimates become less reliable because of significant statistical noise. Mortality of males exceeds mortality of females up to very old ages (110 years). It also was found that life expectancy at age 80 depends on the month of birth: individuals born in January live longer lives than persons born in other months and in April-June in particular. This periodicity repeats in every studied birth cohort starting from the birth years1885 to 1899. However, by age 100 this dependence of survival on month of birth fades out, indicating that centenarians indeed represent a selected population.

This exploratory study has developed a new methodology of using online genealogical, historical and demographic data resources for actuarial longevity studies. It also tested some hypotheses on predictors of human longevity and identified determinants of survival to advanced ages. This exploratory study has demonstrated the feasibility of subsequent large-scale studies on predictors of human longevity, and provided both a preliminary estimate of the magnitude of the effects of these longevity predictors, and a number of new research ideas that can be pursued as testable hypotheses in future actuarial studies.

ACKNOWLEDGEMENTS

This study was made possible thanks to generous support from the Society of Actuaries and a stimulating working environment at the Center on Aging, NORC/University of Chicago. We are grateful to Thomas Edwalds, Kenneth Faig and Ward Kingkade for their helpful comments and suggestions on the initial draft of this manuscript. We also are grateful to the members of the Project Oversight Group (POG) of the Society of Actuaries (Thomas Edwalds, Patricia Pruitt, Elizabeth Arias), and to participants of the international symposium, "Living to 100 and Beyond: Survival at Advanced Ages" (Orlando, Fla., January 12-14, 2005) for providing useful comments and recommendations on the second revised version of this manuscript.

The earlier version of this report (49 pages) was published by the Society of Actuaries in the "Living to 100 and Beyond" monograph in July 2005 (published online at: http://library.soa.org/library-pdf/m-li05-1_V.pdf), and selected by the *North American Actuarial Journal* to be considered for publication there.

BIBLIOGRAPHY

Allison, P. (1995). *Survival Analysis Using the SAS ® System: A Practical Guide*. SAS Institute.

Austad, S.N. (2001). "Concepts and Theories of Aging." In E.J. Masoro and S.N. Austad. *Handbook of the Biology of Aging*. San Diego, CA: Academic Press, 3-22.

Bagui, S. and Earp, R. (2003). *Database Design Using Entity-Relationship Diagrams*. CRC Press: Boca Raton, FL.

Barker, D.J.P. (1998). "Mothers, Babies, and Disease in Later Life." (2nd ed.). London: Churchill Livingstone.

Beard, R.E. (1971). "Some Aspects of Theories of Mortality, Cause of Death Analysis, Forecasting and Stochastic Processes." In W. Brass (Ed.), *Biological Aspects of Demography* (pp. 57-68), London: Taylor and Francis.

Belmont, L. and Marolla, F.A. (1973). "Birth order, family size, and intelligence." *Science* 182: 1096–1101.

Belmont, L., Stein, Z.A. and Wittes, J.T. (1976). "Birth order, family size and school failure." *Developmental Medicine and Child Neurology* 18: 421–430.

Bijur, P.E., Golding, J. and Kurzon, M. (1988). "Childhood accidents, family size and birth order." *Social Science & Medicine* 26: 839–843.

Blake, J. (1981). "Family size and the quality of children." *Demography* 18: 421–442.

Blake, J. (1989). "Number of siblings and educational attainment." *Science* 245: 32–36.

Breland, H.M. (1973). "Birth order, family size, and intelligence." *Science* 184: 149.

Breland, H.M. (1974). "Birth order, family configuration and verbal achievement." *Child Development* 45: 1011–1019.

Bronikowski, A.M., Alberts, S.C., Altmann, J., Packer, C., Carey, K.D. and Tatar, M. (2002). "The Aging Baboon: Comparative Demography in a Non-Human Primate." *Proc. Natl. Acad. Sci. U.S.A.*, 99, 9591-9595.

Burnett, J. (1991). "Housing and the decline of mortality." In: Schofield, R., Reher, D. and Bideau, A., Editors, 1991. *The decline of mortality in Europe*, Clarendon Press, Oxford, pp. 158–176.

Carey, J.R., Liedo, P., Orozco, D. and Vaupel, J.W. (1992). "Slowing of mortality rates at older ages in large medfly cohorts." *Science*, 258, 457-461.

Celik, Y. and Hotchkiss, D.R. (2000). "The socio-economic determinants of maternal health care utilization in Turkey." *Social Science & Medicine* 50: 1797–1806.

Chidambaram, V.C., McDonald, J.W. and Bracher, M.D. (1987). "Childhood mortality." In: Cleland, J. and Scott, C., Editors, 1987. *The world fertility survey: An assessment*, Oxford University Press, Oxford, pp. 862–881.

Costa, D.L. and Lahey, J. (2003). "Becoming Oldest-Old: Evidence from Historical U.S. Data." *NBER Working Paper* No. W9933. http://ssrn.com/abstract=439614.

Curtsinger, J.W., Fukui, H., Townsend, D. and Vaupel, J.W. (1992). "Demography of Genotypes: Failure of the Limited Life-Span Paradigm in *Drosophila melanogaster*." *Science*, 258, 461-463.

Depoid F. (1973). "La Mortalite des Grands Viellards." *Population*, 28: 755-92.

Doblhammer, G. (1999). "Longevity and Month of Birth: Evidence from Austria and Denmark." *Demographic Research* [Online] 1, 1-22. Available at: http://www.demographic-research.org/Volumes/Vol1/3/default.htm

Doblhammer, G. (2003). "The Late Life Legacy of Very Early Life." Rostock, *MPIDR Working Paper WP-2003-030*.

Doblhammer, G., Vaupel J.W. (2001). "Lifespan Depends on Month of Birth." *Proc. Natl. Acad. U.S.A.*, 98: 2934-2939.

Economos, A.C. (1979). "A Non-Gompertzian Paradigm for Mortality Kinetics of Metazoan Animals and Failure Kinetics of Manufactured Products." *AGE*, 2, 74-76.

Economos, A.C. (1980). "Kinetics of Metazoan Mortality." *J. Social Biol. Struct.*, 3, 317-329.

Economos, A.C. (1983). "Rate of Aging, Rate of Dying and the Mechanism of Mortality." *Arch. Gerontol. and Geriatrics*, 1, 3-27.

Edwalds T. (2005). Theories of Longevity [Discussion], pp. 1-6. In: “*Living to 100 and Beyond*" Monograph. The Society of Actuaries, Schaumburg, Illinois, USA. Published online at: http://library.soa.org/library-pdf/m-li05-1_tr4.pdf

Elliot, B.A. (1992). “Birth order and health: Major issues.” *Social Science & Medicine* 35: 443–452.

Elo, I.T. and Preston, S.H. (1992), "Effects of Early-Life Condition on Adult Mortality: A Review." *Population Index*, 58(2): 186-222.

Elo, I.T., Preston, S.H., Rosenwaike, I., Hill, M. and Cheney, T.P. (1995). "Consistency of Age Reporting on Death Certificates and Social Security Administration Records among Elderly African-American Decedents." *Working Paper Series No. 95-03*. Population Aging Research Center, University of Pennsylvania.

Espehaug, B., Daltveit, A.K., Vollset, S.E., Oyen, N., Ericson, A. and Irgens, L.M. (1994). “Infant survival in Norway and Sweden 1985–88.” *Acta Paediatrica* 83: 977–982.

Faig, K. (2001). "Reported Deaths of Centenarians and Near-Centenarians in the U.S. Social Security Administration's Death Master File." In: Proceedings of the Society of Actuaries "Living to 100 and Beyond" International Symposium, Orlando, FL.

Fergusson, D.M., Dimond, M.E., Horwood, L.J. and Shannon, F.T. (1984). “The utilisation of preschool health and education services.” *Social Science & Medicine* 11: 1173–1180.

Fogel, R.W. and Costa, D.L. (1997). "A Theory of Technophysio Evolution, with Some Implications for Forecasting Population, Health Care Costs, and Pension Costs." *Demography*, 34: 49-66.

Gavrilov, L.A. and Gavrilova, N.S. (1991). *The Biology of Life Span: A Quantitative Approach*. Harwood Academic Publisher, New York.

Gavrilov, L.A. and Gavrilova, N.S. (1998). "Inventory of Data Resources on Familial Aggregation of Human Longevity that can be Used in Secondary Analysis in Biodemography of Aging." Bethesda: National Institute on Aging. NIA Professional Service Contract #263 SDN74858. 35p.

Gavrilov, L.A. and Gavrilova, N.S., 1999. "Season of Birth and Human Longevity." *Journal of Anti-Aging Medicine*, 2: 365-366. Available at: http://longevity-science.org/Season-of-Birth.pdf

Gavrilov, L.A. and Gavrilova N.S. (2000). "Human Longevity and Parental Age at Conception." In: *Sex and Longevity: Sexuality, Gender, Reproduction, Parenthood* (J.-M. Robine *et al*., eds), pp. 7-31. Berlin, Heidelberg: Springer-Verlag. Available at: http://longevity-science.org/Parental_Age_2000.pdf

Gavrilov, L.A. and Gavrilova, N.S. (2001a). "Biodemographic Study of Familial Determinants of Human Longevity." *Population, English Selection*, 13(1) 197-222. Available at: http://longevity-science.org/Biodemography-2001.pdf

Gavrilov, L.A. and Gavrilova, N.S. (2001b). "The Reliability Theory of Aging and Longevity." *J. Theor. Biol.* 213: 527-545. Available at: http://longevity-science.org/JTB-01.pdf

Gavrilov, L.A. and Gavrilova, N.S. (2002). "Early-Life Seasonal Programming of Adult Lifespan: Evidence from the 19th Century Birth Cohorts." Annual Meeting of the Social Science History Association, St. Louis, October 24-27, 2002. Available at: http://www.ssha.org/abstract2002/abs348.html

Gavrilov, L.A. and Gavrilova, N.S. (2003a). "The Quest for a General Theory of Aging and Longevity." *Science's* SAGE KE (Science of Aging and Knowledge Environment) for July 16, 2003; Vol. 2003, No. 28, 1-10. Available at: http://longevity-science.org/SAGE-KE-03.pdf

Gavrilov, L.A. and Gavrilova, N.S. (2003b). "Early-Life Factors Modulating Lifespan." In: Rattan, S.I.S. (Ed.). *Modulating Aging and Longevity*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 27-50. Available at: http://longevity-science.org/Early-Life-Factors-2003.pdf

Gavrilov L.A. and Gavrilova, N.S. (2004). "Early-Life Programming of Aging and Longevity: The Idea of High Initial Damage Load (the HIDL Hypothesis)." *Annals of the New York Academy of Sciences*, 1019: 496-501. Available at: http://longevity-science.org/HIDL-ANYAS-2004.pdf

Gavrilov, L.A., Gavrilova, N.S., Olshansky, S.J. and Carnes, B.A. (2002). "Genealogical Data and Biodemography of Human Longevity." *Social Biology*, 49(3-4): 160-173. Available at: http://longevity-science.org/Social-Biology-02.pdf

Gavrilova, N.S. and Gavrilov, L.A. (1999). "Data Resources for Biodemographic Studies on Familial Clustering of Human Longevity." *Demographic Research* [Online], 1(4): 1-48. Available at: http://www.demographic-research.org/volumes/vol1/s4

Gavrilova, N.S. and Gavrilov, L.A. (2001). "When does Human Longevity Start? Demarcation of the Boundaries for Human Longevity." *Journal of Anti-Aging Medicine*, 4**:** 115-124. Available at: http://longevity-science.org/JAAM-Boundaries-for-Human-Longevity.pdf

Gavrilova, N.S., Gavrilov, L.A., Evdokushkina, G.N. and Semyonova, V.G. (2003). "Early-Life predictors of Human Longevity: Analysis of the 19th Century Birth Cohorts." *Annales de Demographie Historique*, 2: 177-198. Available at: http://longevity-science.org/Early-Life-Predictors-2003.pdf

Gompertz, B. (1825). "On the Nature of the Function Expressive of the Law of Human Mortality and on a New Mode of Determining Life Contingencies." *Philos. Trans. Roy. Soc. London* A, 115**:** 513-585.

Greenwood, M. and Irwin, J.O. (1939). "The Biostatistics of Senility." *Hum. Biol.*, 11: 1-23.

Hanushek, E.A. (1992). "The trade-off between child quantity and quality." *Journal of Political Economy* 100: 84–117.

Hill, M.E. and Rosenwaike, I. (2001). "The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages." *Soc. Secur. Bull*: 64: 45-51.

Hill, M.E., Preston, S.H., Elo, I.T. and Rosenwaike, I. (1995). "Age-Linked Institutions and Age Reporting among Older African Americans." Population Research Center, University of Pennsylvania, Working Paper Series No. 95-05.

Hill, M.E., Preston S.H. and Rosenwaike, I. (2000). "Age Reporting among White Americans Aged 85+: Results of a Record Linkage Study." *Demography*, 37: 175-186.

Hill, M.E., Preston, S.H., Rosenwaike, I. and Dunagan, J.F. (2000). "Childhood Conditions Predicting Survival to Advanced Age among White Americans." Paper presented at the 2000 Annual Meeting of the Population Association of America, Los Angeles.

Horiuchi, S. and Wilmoth, J.R. (1998). "Deceleration in the Age Pattern of Mortality at Older Ages." *Demography*, 35: 391-412.

Horwitz, M.S., Morgenstern, H. and Berkman, L.F. (1985). "The impact of social stressors and social networks on pediatric medical care use." *Medical Care* 23: 946–959.

Iwashyna, T.J., Zhang, J.X., Lauderdale, D.S. and Christakis, N.A. (1998). "A Method for Identifying Married Couples in the Medicare Claims Data: Mortality, Morbidity, and Health Care Utilization among the Elderly." *Demography*, 35: 413-419.

Jeune, B. and Vaupel, J., eds. (1999). "Validation of Exceptional Longevity." *Odense Monographs on Population Aging,* Vol. 6. Odense: Odense Univ. Press.

Kannisto, V. (1994). *Development of Oldest-Old Mortality, 1950-1990: Evidence from 28 Developed Countries*. Odense Monographs on Population Aging. Vol. 1. 1994, Odense: Odense University Press.

Kaplan, B.A., Mascie-Taylor, C.G. and Boldsen, J. (1992). "Birth order and the health status in a British national sample." *Journal of Biosocial Science* 24: 25–33.

Kestenbaum, B. (1992). "A Description of the Extreme Aged Population Based on Improved Medicare Enrollment Data." *Demography*, 29: 565-80.

Kestenbaum, B. and Ferguson, B.R. (2001). "Mortality of the Extreme Aged in the United States in the 1990s, Based on Improved Medicare Data." In: Proceedings of the Society of Actuaries "Living to 100 and Beyond" International Symposium, Orlando, FL.

Kestenbaum, B.M. and Ferguson, B.R. (2005). "Number of Centenarians in the United States Jan. 1, 1990, Jan. 1, 2000, and Jan. 1, 2010, Based on Improved Medicare Data." In: "Living to 100 and Beyond" Monograph. The Society of Actuaries, Schaumburg, Illinois, USA. Published online at: http://library.soa.org/library-pdf/m-li05-1_XXVI.pdf

Kuh, D. and Ben-Shlomo, B. (1997). *A Life Course Approach to Chronic Disease Epidemiology*. Oxford: Oxford University Press.

Le Bras, H. (1976). "Lois de Mortalité et Age Limité." *Population*, 31, 655-692.

Leon, D.A., Lithell, H.O., Vågerö, D., Koupilová, I., Mohsen, R., Berglund, L., Lithell, U.B. and McKeigue, P.M. (1998). "Reduced Fetal Growth Rate and Increased Risk of Death from Ischaemic Heart Disease: Cohort Study of 15000 Swedish Men and Women Born 1915-29." *Br. Med. J.* 317: 241-245.

Lindop, P.J. (1961). "Growth Rate, Lifespan and Causes of Death in SAS/4 Mice." *Gerontologia*, 5: 193-208.

Magnus, P., Berg, K. and Bjerkedal, T. (1985). “The association of parity and birth weight: Testing the sensitization hypothesis.” *Early Human Development.* 12: 49–54.

Majumder, A.K. (1988). “Maternal factors and infant and child mortality in Bangladesh.” *Journal of Biosocial Science* 20: 89–98.

Manton, K.G. and Vaupel, J.W. (1995). "Survival After Age of 50 in the United States, Sweden, France, England, and Japan." *The New England Journal of Medicine*. 333: 1232-5.

Modin, B. (2002). "Birth Order and Mortality: A Life-Long Follow-up of 14,200 Boys and Girls Born in early 20th Century Sweden." *Social Science & Medicine*, 54: 1051-1064.

Nixon, J. and Pearn, J. (1978). "An Investigation of Socio-Demographic Factors Surrounding Childhood Drowning Accidents." *Social Science & Medicine* 12: 387–390.

Perks, W. (1932). "On Some Experiments in the Graduation of Mortality Statistics." *Journal of the Institute of Actuaries*, 63, 12-57.

Perls T. (2002). "Genetic and environmental influences on exceptional longevity and the AGE nomogram." *Ann. N. Y. Acad. Sci.* 959: 1-13.

Preston, S.H., Haines M.R. (1991). *Fatal years. Child mortality in late nineteenth-century America*. Princeton, NJ: Princeton University Press.

Preston, S.H., Elo, I., Rosenwaike, I. and Hill, M. (1996). "African-American Mortality at Older Ages: Results of a Matching Study." *Demography,* 33: 193-209.

Preston, S.H., Hill, M. and Drevenstedt, G. (1998). "Childhood Conditions that Predict Survival to Advanced Ages among African Americans." *Soc. Sci. Med.*, 47: 1231-1246.

Rosenwaike, I. and Logue, B. (1983). "Accuracy of Death Certificate Ages for the Extreme Aged." *Demography*, 20: 569-85.

Rosenwaike, I., Hill, M., Preston, S. and Elo, I. (1998). **"**Linking Death Certificates to Early Census Records: The African American Matched Records Sample." *Historical Methods*, 31: 65-74.

Rosenwaike, I. and Stone, L.F. (2003). "Verification of the Ages of Supercentenarians in the United States: Results of a Matching Study." *Demography,* 40: 727-739.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C.A., Goeken, R., Hall, P.K., King, M. and Ronnander, C. (2004). *Integrated Public Use Microdata Series (IPUMS): Version 3.0.* Minneapolis, MN: Minnesota Population Center. Available at: http://www.ipums.org.

Sacher, G.A. (1966). "The Gompertz transformation in the Study of the Injury-Mortality Relationship: Application to Late Radiation Effects and Ageing." In P. J. Lindop and G. A. Sacher (eds.) *Radiation and Ageing* (pp. 411-441). Taylor and Francis, London.

Sears, R.R., Maccoby, E.E. and Levin, H. (1957). *Patterns of child rearing*, Row Peterson, Evanston, IL.

Shrestha, L.B. and Preston, S. H. (1995). "Consistency of Census and Vital Registration Data on Older Americans: 1970-1990." *Survey Methodology*, 21: 167-177.

Stata Corp. (2005). *Stata User's Guide.* Stata Press. http://www.stata.com.

Thatcher, A.R. (1999). "The Long-Term Pattern of Adult Mortality and the Highest Attained Age." *J. R. Statist. Soc.* A, 162, Part 1, 5-43.

Vaupel, J.W., Carey, J.R., Christensen, K., Johnson, T., Yashin, A.I., Holm, N.V., Iachine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.D., Zeng, Y., Manton, K. and Curtsinger, J.W. (1998). "Biodemographic Trajectories of Longevity." *Science*, 280, 855-860.

Vincent, P. (1951). "La mortalite des Viellards." *Population*, 6, 181-204.

APPENDIX I

Genealogical and historical resources used in this project

The resources were provided by the MyFamily.com, Inc. (Provo, Utah) network of genealogical resources. MyFamily brings together the most comprehensive collection of services to help researchers in family reconstitution and linkage to official documentary sources.

The following MyFamily services were used in this project:

Ancestry.com

URL: http://ancestry.com

This service provides online access to the following historical resources:

1. Collection of U.S. censuses from 1790 to 1930. This collection contains indexes for enumerated individuals (1900, 1910, 1920, 1930) or heads of household (other censuses). Allows search by first and last names, state, county, relation to head of household, year of birth across over 75 million individuals. Microfilm images of the United States federal censuses also are available online for subscribers.

2. Social Security Administration Death Master File. Records in this database are linked to the search engine, which allows researchers to conduct search in other databases including online U.S. censuses.

3. Obituary collection. This is a collection of images taken from local newspapers as well as a collection of tombstone inscriptions.

4. State collections of birth and death records including California Death Index, 1940-1997 and Texas Deaths, 1964-1998.

Ancestry.com has many other collections including collections of local newspapers, World War I Draft Registration Cards, 1917-1918, Family and local history collections, U.K. census collections, U.S. immigration collections and others. In this project we used the collections listed above (numbers 1-4) and collections of computerized genealogies called the OneWorldFamily collection.

Genealogy.com

URL: http://genealogy.com

This service has a collection of U.S. censuses from 1790 to 1930. 1900 and 1910 censuses are indexed by head of household. The main advantage of this collection is the very high quality of images of census pages. The search capabilities of this service are less powerful compared to Ancestry.com, so the best research strategy is to locate the census page using Ancestry.com and then download an image of a previously located census page using the Genealogy.com service.

Rootsweb.com

URL: http://www.rootsweb.com

This is a free service, which contains a free version of the Social Security Death Master File and a collection of computerized genealogies submitted by volunteers (named the WorldConnect project).