SEARCH FOR PREDICTORS OF EXCEPTIONAL HUMAN LONGEVITY: USING COMPUTERIZED GENEALOGIES AND INTERNET RESOURCES FOR HUMAN LONGEVITY STUDIES

Natalia S. Gavrilova* and Leonid A. Gavrilov†

ABSTRACT

This paper explores new opportunities provided by the ongoing revolution in information technology, computer science, and Internet expansion for studies of exceptional human longevity. To this aim, the detailed family data for 991 alleged centenarians born between 1875 and 1899 in the United States were extracted from publicly available computerized family histories of 75 million individuals available at the Rootsweb site. To validate the age of the centenarians, these records were linked first to the Social Security Administration Death Master File records (for death date validation) and then to the records of the U.S. censuses for 1900, 1910, and 1920 (for birth date validation). The results of this cross-validation study demonstrated that computerized genealogies may serve as a useful starting point for developing a reliable family-linked scientific database on exceptional human longevity.

The resulting database on centenarians with validated ages was used in the study of the predictors of exceptional human longevity, including familial factors and early-life living conditions. 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 for the 1900 U.S. census suggests that a farm background (farm ownership by parents in particular) and childhood residence in the Western region of 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.

Analysis of familial factors suggests that there may be a link between exceptional longevity and a person's birth order. It was 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. Further within-family comparison of centenarians with their siblings found that the protective effect of being first-born is driven mostly by the young maternal age at the person's birth (being born to a mother younger than 25 years). Being born to a young mother is an important within-family predictor of human longevity, and even at age 75 it is still important to be born to young mother to survive to 100 years.

* Natalia Gavrilova, PhD, is a research associate in the Center on Aging of the National Opinion Research Center at the University of Chicago, 1155 E. 60th St., Chicago, IL 60637, gavrilova@longevity-science.org.

† Leonid Gavrilov, PhD, is a research associate in the Center on Aging of the National Opinion Research Center at the University of Chicago, 1155 E. 60th St., Chicago, IL 60637, gavrilov@longevity-science.org.

1. 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 and demographic studies. The number of U.S. centenarians is growing at a rate of about 4.1% per year, so their numbers have increased by 51% in the 10-year period from January 1, 1990, to January 1, 2000 (Kestenbaum and Ferguson 2005). Yet factors predicting exceptional longevity and its time trends remain to be fully understood.

Our previous search for additional data resources (see Gavrilov and Gavrilova 1998; Gavrilova and Gavrilov 1999) revealed an enormous amount of new lifespan data that could be made readily available for subsequent full-scale studies. Millions of genealogical records are already computerized and after strict data validation could be used for the study of familial and other predictors of human longevity. 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 ancestors. The compilers of genealogies aided this time-consuming task by using many different sources: genealogical libraries, The Church of Jesus Christ of 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, when compared to other data sources, such as death certificates, census data, and the Medicare database. For example, 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 (SSA) NUMIDENT file contains information on the names of parent-child pairs for Medicare beneficiaries (65 years and older). In the latter case, however, only information for persons surviving to 65 may be obtained. Thus, when

compared to other resources, computerized genealogies provide unique opportunities to greatly expand the scope of actuarial studies of human longevity, and other data resources cannot adequately be substituted for them. However, one substantial limitation of genealogical data is their uncertain quality, which requires additional efforts of data quality control before using this resource in scientific research.

In this study we describe our experience in identification, collection, verification, and analysis of data taken from computerized genealogies for long-lived individuals born in the United States. 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 (Elo and Preston 1992; Gavrilov and Gavrilova 1991, 2003a; Barker 1998; Kuh and Ben-Shlomo 1997; Costa and 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 and 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 and Gavrilova 1991, 2001, 2006). 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 influencing the level of this damage (Gavrilov and Gavrilova 1991, 2001, 2003a, 2004, 2006). This prediction appears to be confirmed for such early-life indicators as the parental age at a person's conception (Gavrilov and Gavrilova 1997, 2000, 2003b; Gavrilova et al. 2003) and the month of a person's birth (Gavrilov and Gavrilova 1999, 2003b; Gavrilova et al. 2003; Doblhammer 1999; Doblhammer and Vaupel 2001; Costa and Lahey 2003). There is mounting evidence now in support of the idea of fetal ori-

gins of adult degenerative diseases (Barker 1998; Kuh and Ben-Shlomo 1997) and early-life programming of aging and longevity (Gavrilov and Gavrilova 1991, 2001, 2003b; Gavrilova et al. 2003). Some of these ideas and predictors were tested in this study.

The results presented in this article 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.

2. COMPUTERIZED GENEALOGIES AS A DATA RESOURCE FOR ACTUARIAL STUDIES OF MORTALITY

2.1 Survey of the Existing Computerized Genealogies

At the first stage of the study, 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 identified in our previous studies (Gavrilov and Gavrilova 1998; Gavrilova and Gavrilov 1999). Centenarian family histories were drawn from computerized family trees using the following selection criteria: (1) persons should have both the birth and death date information and have lifespans of 100 years and over, (2) persons should be born in the United States after 1875, and (3) 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 the 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 likely immigrated 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 study is focused on the role of childhood condi-

tions 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 the early U.S. censuses. It should also be noted that foreign-born children made up a small proportion (3%) of all children below age 10 enumerated in the 1900 census according to the estimate obtained from the Integrated Public Use Microdata Series (IPUMS) 1% random sample of the 1900 U.S. census population (Ruggles et al. 2004).

Using online genealogical data resources, we identified over 2000 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 and Gavrilov 1999). GEDCOM stands for the *Genealogical Data Communication* standard proposed by the Family History Department of The Church of Jesus Christ of Latter-Day Saints 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.

2.2 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: individuals and unions/marriages with one-to-many relationships (an individual 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 Death Master File 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 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 2004 long-lived individuals and their relatives (including parents and grandparents) in the form of a relational database. From these 2004 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 had an uncertain quality, which required 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, Preston, and Rosenwaike 2000; Rosenwaike and Stone 2003; Shrestha and 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.

2.3 Verification of Centenarian Birth and Death Dates

Verification of centenarian birth and death dates was made in three steps. To check for obvious mistakes in 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.

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

The verification of death dates is an important step in data 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 Death Master File (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. This resource covers deaths that occurred in the period 1937–2005 (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, whereas age reporting in death certificates is made by a proxy informant (Kestenbaum 1992; Kestenbaum and Ferguson 2001; Rosenwaike et al. 1998; Rosenwaike and Stone 2003). Therefore, we based the death date verification on linkage to the DMF, which is publicly available at the Rootsweb web site (Faig 2001).

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 of age and number of months not employed; information about school attendance and literacy; and information about home ownership or farm residence.

In our study the linkage of centenarian records to the early census data was facilitated by online availability of the entire indexed U.S. 1900, 1910, and 1920 censuses, a service provided by Genealogy.com and Ancestry.com. In this study we conducted a linkage of 534 centenarian records (for centenarians found in the 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.

2.4 Description of the Verified Data Sample of Centenarians and Quality of Centenarian Genealogical Data

The data consistency checks revealed a surprisingly small number of obvious data inconsistencies in computerized genealogies. In one case the alleged centenarian had parents with incorrect birth dates (parents born later than the person himself). This case was dropped from the study. 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 DMF but not found in early censuses and therefore not included in the final analyses). All other records did not reveal obvious inconsistencies in event dates, so

990 records were left for further verification. A general overview of the data collection, verification, and linkage used in this study is presented in Figure 1.
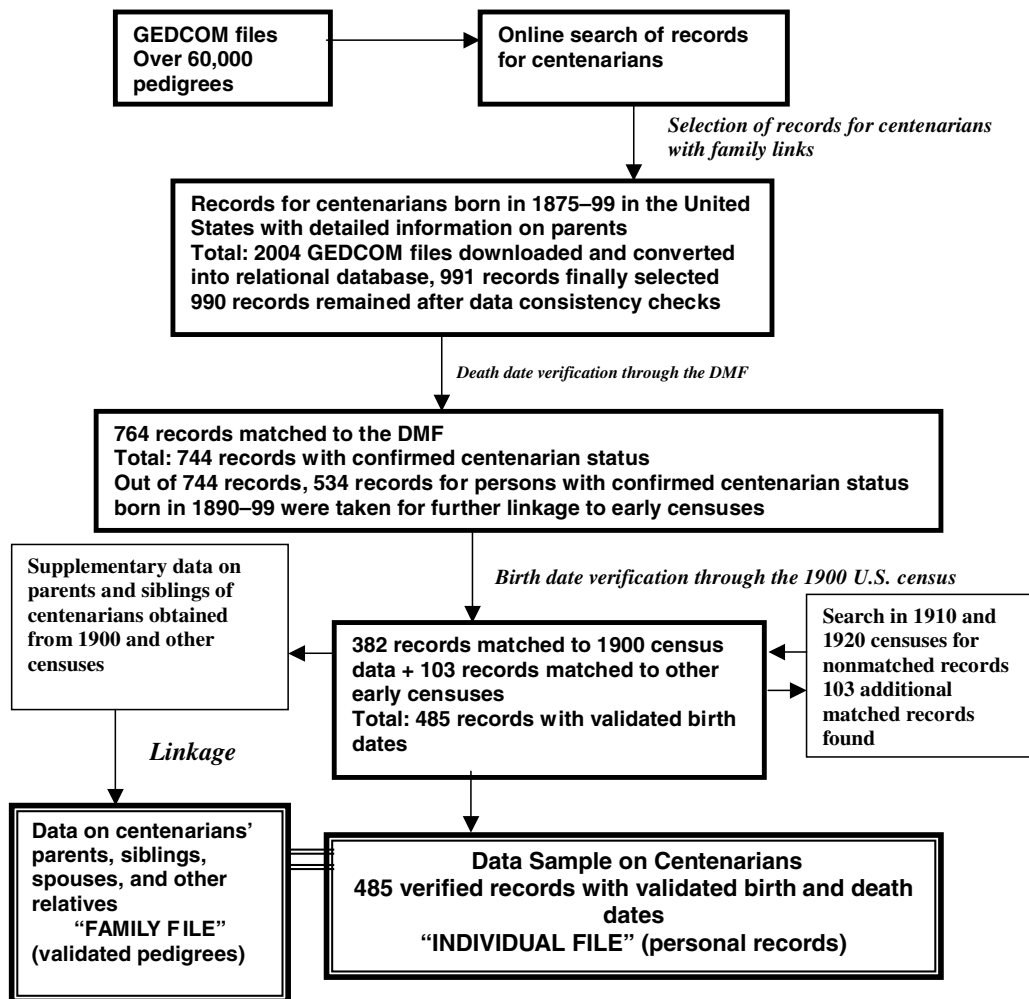
The overall linkage success rate to the SSA DMF was moderate at 75% for males and 78% for females. Among the 767 persons found in the DMF, centenarian status could not be confirmed for only 23 alleged centenarians from the computerized genealogies (3.0%). The detailed breakdown of records found in the DMF is presented in Table 1. In 17 cases (2%) the difference between the death year in the genealogy, and the DMF was expressed in round numbers (e.g., 10, 20, or 30 years), which seems to be caused by misprints in genealogies (see Table 1). Thus all cases of exceptional longevity in genealogies should be checked using the SSA or vital statistics records.

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. Also, the DMF covers about 90% of all deaths for which death certificates are issued (see Faig 2001) and about 92–96% of deaths for persons older than 65 years (Hill and Rosenwaike 2001). Further work with nonmatched cases using additional data sources (obituaries, state collections of death certificates) revealed that about half of nonmatched cases are related to misprints in genealogies, and about 20% of nonmatched cases have correct death dates (as confirmed by linkage to the state death indexes) although not recorded in the DMF.

It should be noted that the linkage success rate to the DMF was substantially higher for persons born after 1889, at 82%. This result is consistent with previous reports that quality and coverage of the DMF database were lower for persons born before 1890 (Faig 2001). 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.

The overall success rate for linkage of centenarian records to early U.S. censuses was 91% (485 out of 534 records; see also Table 2). There was no significant difference between individuals who were successfully linked to early censuses

Figure 1
General Overview of Data Collection and Data-Processing Protocol

```
┌──────────────────┐          ┌──────────────────┐
│ GEDCOM files     │─────────▶│ Online search of │
│ Over 60,000      │          │ records          │
│ pedigrees        │          │ for centenarians │
└──────────────────┘          └──────────────────┘
```

Selection of records for centenarians with family links

```
┌──────────────────────────────────────────────────────┐
│ Records for centenarians born in 1875–99 in the United│
│ States with detailed information on parents            │
│ Total: 2004 GEDCOM files downloaded and converted      │
│ into relational database, 991 records finally selected │
│ 990 records remained after data consistency checks     │
└──────────────────────────────────────────────────────┘
```

Death date verification through the DMF

```
┌──────────────────────────────────────────────────────┐
│ 764 records matched to the DMF                         │
│ Total: 744 records with confirmed centenarian status   │
│ Out of 744 records, 534 records for persons with       │
│ confirmed centenarian status born in 1890–99 were taken│
│ for further linkage to early censuses                  │
└──────────────────────────────────────────────────────┘
```

Birth date verification through the 1900 U.S. census

```
┌───────────────────┐    ┌──────────────────────┐    ┌──────────────────┐
│ Supplementary data│    │ 382 records matched  │    │ Search in 1910 and│
│ on parents and    │◀───│ to 1900 census data +│◀───│ 1920 censuses for │
│ siblings of       │    │ 103 records matched  │───▶│ nonmatched records│
│ centenarians      │    │ to other early       │    │ 103 additional    │
│ obtained from 1900│    │ censuses             │    │ matched records   │
│ and other censuses│    │ Total: 485 records   │    │ found             │
└───────────────────┘    │ with validated birth │    └──────────────────┘
                         │ dates                │
                         └──────────────────────┘
```

Linkage

```
┌────────────────────┐    ┌──────────────────────────────────────┐
│ Data on            │    │ Data Sample on Centenarians           │
│ centenarians'      │    │ 485 verified records with validated   │
│ parents, siblings, │    │ birth and death dates                 │
│ spouses, and other │    │ "INDIVIDUAL FILE" (personal records)  │
│ relatives          │    └──────────────────────────────────────┘
│ "FAMILY FILE"      │
│ (validated         │
│ pedigrees)         │
└────────────────────┘
```

Notes: 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.

and nonlinked ones regarding their birth year, birth order, marital status, and sibship size (although the fathers of nonlinked individuals died before age 50 slightly more often).

The agreement between dates of birth recorded in computerized genealogies and dates 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% of cases. In one case only, the centenarian's year of birth was three years less than in the genealogy record, that is, the centenarian was in fact *older* than was reported in the genealogy file. In 4.5% of cases, 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% of cases the centenarian was one year younger than reported in genealogy. Disagreements between birth years reported in early censuses and genealogies were more notable for parents (about 15% of all cases) than for children, but in the majority of cases the differences did not exceed one year. 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 3.

Thus, we obtained 485 records for centenarians with verified birth dates, confirmed centenarian

Table 1
Comparison of Death Year Reporting in Genealogical Records and the Social Security Administration Death Master File (SSA DMF)

Age at Death Reported in Genealogy	Difference between the Death Year Reported in Genealogy and SSA DMF	Number of Cases
100	−1	1
	0	216
	1	1
	10	6
	20	1
101	−1	4
	0	241
	1	4
	10	2
	20	3
102	−1	2
	0	154
	2	1
	20	1
103	0	73
	1	1
	22	1
104	0	23
105	0	9
106	0	12
	20	2
107	0	5
	17	1
109	0	1
110	30	1
114	30	1

Table 3
Summary of Results of Genealogy Records' Linkage First to the 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%)

provide a detailed date of birth (day, month, and year) most likely taken from birth certificates or family Bible records, whereas census records are based on verbal reports during enumeration. This study demonstrated the feasibility of large-scale data linkage to historical online resources. With the help of online indexes, linkage to early censuses is no longer as laborious and time-consuming procedure as it was in the past.

As a result of this validation study, a sample of 485 centenarians born in the United States in 1890–1900 was identified. All centenarians had verified dates of birth and death and known information for parents, siblings, spouses, and other relatives. Table 4 shows the age and sex breakdown of centenarians with verified ages.

The database on long-lived persons developed in this study combines information on family characteristics with data on the early-life conditions taken from the 1900–1910 U.S. censuses. This database was used in testing several hypotheses on the factors affecting exceptional longevity. The statistical analyses were conducted using Stata (Stata Corp, 2005).

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 studied 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

Table 2
Number and Percentage of Genealogical Records Successfully Linked to Early U.S. Census Records (for Records Already Confirmed through Linking to the 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

Table 4
Distribution of Centenarians Confirmed through Linkage to the DMF 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

3. STUDY OF EARLY-LIFE CHILDHOOD CONDITIONS AND LONGEVITY

3.1 Using the U.S. Census of Population Data to Study Early-Life Predictors of Longevity

The resulting data set 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, Hill, and Drevenstedt (1998) and used individual data from the 1900 U.S. census of population as a control group. The data are available as a part of the IPUMS project at the University of Minnesota (Ruggles et al. 2004). The sample represents 1% of Caucasian households enumerated in 1900 (households where the head of household was Caucasian). The linkage to early U.S. censuses in our study found that most centenarians in our sample were Caucasian (with the exception of two Native American families), so we used a Caucasian population sample from the IPUMS data set as a control. At this initial stage of the data analysis, we conducted a comparison of households that raised a future centenarian and were linked to the 1900 census (358 cases) to the population-based sample of Caucasian households enumerated by the 1900 census, which had children below age 10, to make these households comparable to the set of centenarians who were born in 1890–99 and hence were below age 10 in 1900 (31,322 control households).

We applied a method of multiple logistic regression (procedure "logistic" in the Stata statistical package) to compare these two sets of households. The variables used for describing a household are similar to those applied by Preston, Hill, and Drevenstedt (1998). We did not use a 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 classification of diverse occupations. In fact, 63% 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.

Our tested hypothesis is that if early childhood conditions are important for survival to age 100, then the households with child future centenarians would be different from the general population. Tables 5 and 6 present results from multivariate logistic regression that estimate the odds for the household to be in the "centenarian" group. We conducted statistical analyses separately for male and female centenarians because our previous studies demonstrated that men and women may respond differently to early-life living conditions (Gavrilov and Gavrilova 1999, 2003b).

Data presented in Tables 5 and 6 demonstrate that the region of childhood residence and the household property status are the two most significant variables that affect chances of a household to produce a future centenarian (for both sons and daughters). Thus, spending a childhood in the Mountain Pacific and West Pacific regions in the United States may highly increase chances of long life (by a factor of three) compared to the Northeastern part of the country. Also a farm (particularly an owned farm) residence is associated with better survival to advanced ages. This result is consistent with the studies of childhood conditions and survival to age 85+ (Preston, Hill, and Drevenstedt 1998; Hill et al. 2000). These earlier studies, also based on linkage to early censuses, demonstrated a significant advantage in survival to age 85 for children living on farms for both African Americans (Preston, Hill, and Drevenstedt 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). These earlier mentioned studies of childhood conditions and later survival also found that a father's illit-

Table 5

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 6

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

eracy significantly decreases the chances of survival to age 85+. We found no such relationship for survival to age 100 in our data set.

Having an immigrant father decreases one's chances to become a centenarian for both 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). 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 of survival to age 85 among African Americans (Preston, Hill, and Drevenstedt 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. Overall, our results are generally consistent with findings obtained in previous studies.

3.2 Links between Birth Order and Exceptional Longevity

Information about birth order of centenarians allowed us to test 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 (random uniform distribution for cases of exceptional longevity by birth order), then the ratio of [centenarian birth order/(sibship 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 should demonstrate a statistically significant deviation from the expected value of 0.5.

To study the birth-order effects, we have to remove noninformative 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). We found that the centenarian birth-order ratio for female centenarians is 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% lower on average than would be expected by pure chance. Thus, it is less likely to find female centenarians 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 (see, however, later results and discussion).

Similar results are obtained using another statistic named "centenarian birth order difference": [centenarian birth order − (sibship 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. We found that the mean value of the centenarian birth-order difference for females is lower (−0.50 ± 0.11) than zero ($P < 0.01$). Thus, there is a tendency for female centenarians to have a smaller birth order (by 0.5 birth-order rank on average) when compared to other siblings. In contrast to centenarian women, the birth-order difference for centenarian males (−0.13 ± 0.22) is close to the theoretically predicted zero value.

The results we presented so far are based on summary statistics, which describe the overall shift in birth-order ranking of centenarians relative to their siblings. Further analyses of the odds of becoming a centenarian as a function of birth order (with centenarian siblings born in the same time window [1890–1900] used as a control group) found that the best fit of the data for both males and females analyzed separately could be achieved with the following polynomial logistic model:

Logit (Longevity odds ratio)

$$= a\,x + b\,x^2 + c\,z + d,$$

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 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 for both 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 a graphical 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).

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 Fig. 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 Fig. 2). However, the 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.

3.3 Within-Family Analysis of Exceptional Longevity: Effects of Birth Order and Parental Age

To test whether the survival advantage of first-born children is indeed a true within-family phenomenon, rather than an artifact of mixing different families together, a method of within-family analysis has been applied to investigate the occurrence patterns for centenarians among siblings, which allows researchers to avoid confounding caused by between-family variation. To conduct these analyses we selected 198 validated centenarians born in 1890–93 (extracted from our centenarian database) and reconstructed their complete family histories using the U.S. censuses, SSA data, genealogical records, and other supplementary data resources. All birth dates of centenarian siblings were reconstructed using information available in computerized genealogies and early censuses. Death dates were reconstructed for 85% of siblings using the DMF, state death indexes, and online genealogies. Birth order of all allegedly first-born centenarians was verified using the data from early censuses. Statistical analyses were conducted applying a method of within-family comparison known as a

Figure 2

Dependence of the Odds of Becoming a Centenarian on Person's Birth Order as Predicted by the Fitted Polynomial Logistic Model
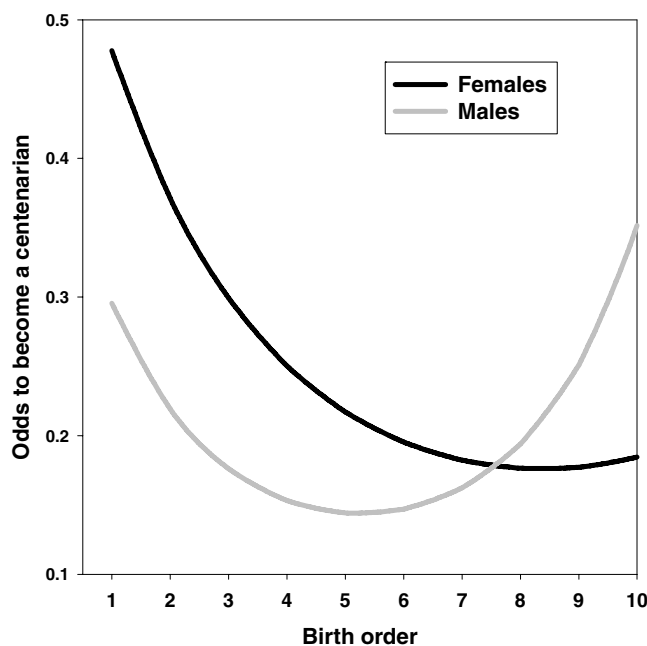
conditional logistic regression (procedure ''clogit'' in Stata). The method of conditional logistic regression allowed us to compare centenarians with their siblings within the same family. This eliminates confounding caused by between-family variation. Using this method we found that the odds of becoming a centenarian are indeed 1.7 times higher for first-born children compared to their later-born siblings (brothers and sisters) from exactly the same family (see Table 7, Model 1).

The next question explored in this study was about the role of child mortality, which was very high a century ago, when the studied centenarians were born. So if first-born children were more likely to survive to an adult age, then simply because of this selective child survival, the centenarians might become more prevalent among the first-born. To test this hypothesis we reanalyzed the data including only those siblings who survived to adulthood (20 years old and above). It was found that even for adult persons the odds of

living to 100 are almost twice higher for first-born persons (Table 7, Model 2). Moreover, even at age 75 it still helps to be a first-born person: the odds of celebrating the 100th birthday are 1.6 times higher for first-born rather than for later-born persons in the same family (Table 7, Model 3).

Then we explored the role of the father's age as a potential explanation for the birth-order effect. When the first child is born, the father is younger and can provide resources for this child for a longer period of time than for his later-born children. Therefore, we tested a hypothesis that it could be the younger age of the father that is responsible for the first-order effect. The collected data on father's age at birth for each studied person were included in the statistical analysis. It turned out that the young father's age was far less important than the first-born status itself in predicting the chances of exceptional longevity. Thus, the hypothesis of a young father as an explanation of the survival advantage of first-born children was rejected (Table 7, Model 4).

Table 7

Odds of Becoming a Centenarian as Predicted by Conditional Logistic Regression (Fixed Effects)

Variable	Odds Ratio	*P* Value	95% Confidence Interval
Model 1: First-born status and sex Number of observations: 950; LR $\chi^2 = 33.75$; Prob. $> \chi^2 = 0.0000$			
First-born status Male sex	1.772 0.404	0.006 0.000	1.180–2.663 0.284–0.576
Model 2: First-born status and sex; survivors to age 20 Number of observations: 797; LR $\chi^2 = 27.54$; Prob. $> \chi^2 = 0.0000$			
First-born status Male sex	1.949 0.458	0.003 0.000	1.261–3.010 0.318–0.658
Model 3: First-born status and sex; survivors to age 75 Number of observations: 557; LR $\chi^2 = 19.03$; Prob. $> \chi^2 = 0.0001$			
First-born status Male sex	1.659 0.459	0.040 0.000	1.022–2.693 0.306–0.687
Model 4: First-born status, paternal age, and sex Number of observations: 950; LR $\chi^2 = 34.24$; Prob. $> \chi^2 = 0.0000$			
First-born status Born to young father ($<$ 25) Male sex	1.635 1.294 0.407	0.039 0.484 0.000	1.025–2.607 0.628–2.668 0.285–0.580
Model 5: First-born status, maternal age, and sex Number of observations: 950; LR $\chi^2 = 39.05$; Prob. $> \chi^2 = 0.0000$			
First-born status Born to young mother ($<$ 25) Male sex	1.360 1.760 0.407	0.189 0.021 0.000	0.859–2.153 1.089–2.846 0.285–0.580
Model 6: Maternal age and sex; survivors to age 75 Number of observations: 557; LR $\chi^2 = 21.31$; Prob. $> \chi^2 = 0.0000$			
Born to young mother ($<$ 25) Male sex	1.869 0.461	0.012 0.000	1.145–3.051 0.307–0.690

Finally, we included into our analysis the mother's age, and it turned out that the young maternal age at childbirth was the most important predictor of exceptional survival, while the effect of the birth order itself has become statistically insignificant. These findings indicate that the beneficial effect of being first-born is driven mostly by the young maternal age at a person's birth (being born to a mother younger than 25 years old). Being born to a young mother is the major predictor of human longevity with the odds ratio to living to 100 being 1.8 times higher than for later-born children, even when the effects of birth order are taken into account (Table 7, Model 5).

Moreover, even at age 75 it is still important to be born to a young mother to survive to 100 years, because the odds of exceptional survival are 1.9 times higher than for later-born siblings (Table 7, Model 6). What is really interesting is that the survival benefits of being born to a young mother are observed only when the mother is younger than 25 years old (data not shown).

Thus, within-family analysis of the birth-order effects on human longevity revealed that it is the young age of mother that is responsible for beneficial effects of first-born status on longevity. A within-family approach has great advantages over other methods, because it is free of confounding caused by between-family differences. However, it remains to be seen whether the observed effect could be reproduced in further studies.

4. Discussion

Computerized genealogies contain important information about family and life-course event that 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 nonblood 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 often are uncertain, resulting in underutilization of this data resource by researchers. In this study we developed a technique of genealogical data collection, verification, and utilization in the scientific analyses of longevity.

This study demonstrated that the ongoing revolution in information technologies has created unique opportunities for conducting actuarial studies of childhood predictors of exceptional longevity. 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%, which is significantly higher than in other studies on the linkage to early censuses: 39–56% (Rosenwaike and Logue 1983; Guest 1987; Rosenwaike et al. 1998), 69% (Hill, Preston, and Rosenwaike 2000), 54% overall, and 69% for Caucasians (Rosenwaike and Stone 2003).

The reasons for this relatively high success rate of linkage to the early censuses in our study can be explained by 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 or 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, Preston, and Rosenwaike 2000; Rosenwaike and Stone 2003).

We had no need to apply the scoring system of match rating suggested in previous studies (Hill, Preston, and Rosenwaike 2000; Rosenwaike and Stone 2003), because the availability of supplemental information in genealogy records made the judgment about match or nonmatch perfectly clear. If the names and years of birth for parents and siblings are in good agreement in both the genealogy 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 the 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 or 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 old, whereas his namesake—his son—died 48 years later (Jeune and 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%) reported all children, including those who died in infancy. Only in two out of 198 such families was a younger child 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, nonreporting a deceased child in a genealogy, and survival of a sibling to an advanced age (even the 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 the quality of preselected 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 a genealogy, then such genealogies might be considered to be a good starting point for further studies. We found that the quality of reporting of birth dates in genealogies is particularly high. Frequency of serious misprints in death dates is higher, although even in this case it is close to 2% 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 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. 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 U-shaped (see Fig. 2).

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, Maccoby, and Levin 1957; Nixon and Pearn 1978; Kaplan, Mascie-Taylor, and Boldsen 1992; Elliot 1992) and educational achievement (Belmont and Marolla 1973; Belmont, Stein, and Wittes 1976; Breland 1973, 1974).

Interestingly, a 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" (Modin 2002, p. 1059), whereas 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 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" (Modin 2002, pp. 1051–52). 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" (Modin, 2002 p. 1059). Later-born children are also heavier at birth, on average, which is considered to provide a survival advantage (Magnus, Berg, and Bjerkedal 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). It 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).

Thus, the initial finding that the first-born children are more likely to become centenarians agreed with the results of other studies, but the mechanism of this birth-order effect remained unclear. A more thorough analysis using within-family analysis showed that it is the young age of the mother that explains the birth-order effect. This may have important social implications, because so many women now decide to postpone childbearing due to career demands. Why being born to a particularly young mother is so beneficial for long-term survival is not yet known, and we plan to test a number of competing social and biological explanations in the future. For example, if the best (most vigorous) maternal ova cells are used first, for the very early pregnancies, this could explain why particularly young mothers produce particularly longevous children. This hypothesis is supported by observations that the first oocytes to have ceased proliferation and entered meiosis are the first to ovulate in the young woman. The last to cease proliferating and arrest

at meiosis ovulate only late in the life of the woman (Keefe, Marquard, and Liu 2006). There is also another hypothesis that may explain the observed findings. It may be reasonable to assume that some particularly young women may be initially free of many diseases and latent infections interfering with optimal fetus development, but later the majority of women become carriers of latent pathogens and conditions simply because of accumulated lifetime exposure. Further testing of these and other related hypotheses is very important and may bring new fresh approaches for enhancing human health and longevity.

Data from early censuses linked to computerized genealogies add additional important information about living 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 general, our results support the idea that early childhood conditions might be important for survival to advanced ages (Gavrilova et al. 2003; Costa and Lahey 2003). Possible mechanisms of these early-life effects were discussed at the international symposium "Living to 100 and Beyond," where an earlier version of this study was presented. In particular, Thomas Edwalds, president of the Chicago Actuarial Association, and the head of Mortality Research at Munich American Reassurance Company, made the following useful suggestions, 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" (Edwalds 2005, p. 5).

Indeed, our findings are consistent with the hypothesis 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 conjecture follows from the comparison of 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 (Preston and Haines 1991, p. 113) and more long-lived children (our study, Tables 5 and 6) 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 (Preston and Haines 1991, p. 112). 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 5 and 6).

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 (Preston and Haines 1991, p. 114). 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 (Preston and Haines 1991, p. 114). 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.

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 above). Our findings agree with the previous reports on the effects of childhood conditions on survival to advanced ages (Preston, Hill, and Drevenstedt 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 population sample. One potential bias from the random sample is a low proportion of minorities in computerized family histories as a result of a requirement to have exact birth and death dates. This requirement may select against family histories of African Americans, who are less likely to report exact birth dates because of lower interaction with age-linked institutions in the past (Hill et al. 1995). For other studied variables, the possibility of bias is not so obvious. 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 documentary trail 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 Caucasian population in general. The definite answer to this question could be obtained by comparing computerized genealogies for ''normal'' (noncentenarian) individuals with population characteristics drawn from the IPUMS database, a study that we hope will be conducted.

5. Conclusions

This study has a number of 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 available online (SSA DMF, 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 U.S. census, for 1930.

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 study we found that the best way to start human longevity database development is to use family-linked data available in computerized genealogies.

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

1. To use only those genealogical records that contain complete, exact, and detailed information on dates of birth and death, places of birth, parental names, and lifespans
2. To use these genealogical data as a starting point only, subject to subsequent external validation with the DMF, early U.S. censuses, and other independent data resources.

Perhaps most important, a particular procedure of data matching and cross-checking has been applied in practice, which produced a reliable data set with several hundred family-linked records for individuals with exceptional longevity. 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 birthplace (or factors associated with it) within the United States seems to be an important determinant of human longevity. Our findings suggest that there may be a threefold difference in the 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 (such as 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.

This study has developed a methodology of using online genealogical, historical, and demo-graphic 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 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.

The results of this study demonstrate that childhood conditions may be indeed very important in determining the chances of exceptional longevity and justify the feasibility of the subsequent large-scale research efforts in this direction.

6. ACKNOWLEDGMENTS

REFERENCES

BAGUI, SIKHA, AND RICHARD EARP. 2003. *Database Design Using Entity-Relationship Diagrams*. Boca Raton, FL: CRC Press.

BARKER, DAVID J. P. 1998. *Mothers, Babies, and Disease in Later Life*. 2nd ed. London: Churchill Livingstone.

BELMONT LILLIAN, AND FRANCIS A. MAROLLA. 1973. Birth Order, Family Size, and Intelligence. *Science* 182: 1096–1101.

BELMONT, LILLIAN, ZENA A. STEIN, AND JANET T. WITTES. 1976. Birth Order, Family Size and School Failure. *Developmental Medicine and Child Neurology* 18: 421–30.

BRELAND, HUNTER M. 1973. Birth Order, Family Size, and Intelligence. *Science* 184: 149.

———. 1974. Birth Order, Family Configuration and Verbal Achievement. *Child Development* 45: 1011–19.

COSTA, DORA L., AND JOANNA LAHEY. 2003. Becoming Oldest-Old: Evidence from Historical U.S. Data. NBER Working Paper No. W9933. http://ssrn.com/abstract=439614.

DOBLHAMMER, GABRIELE. 1999. Longevity and Month of Birth: Evidence from Austria and Denmark. *Demographic Research* [Online] 1: 1–22. www.demographic-research.org/volumes/vol1/3/.

DOBLHAMMER, GABRIELE, AND JAMES W. VAUPEL. 2001. Lifespan Depends on Month of Birth. *Proceedings of the National Academy of Sciences USA* 98: 2934–39.

EDWALDS, THOMAS. 2005. Theories of Longevity [Discussion]. In *Living to 100 and Beyond,* pp. 1–6. http://library.soa.org/library-pdf/m-li05-1_tr4.pdf.

ELLIOT, BARBARA A. 1992. Birth Order and Health: Major Issues. *Social Science & Medicine* 35: 443–52.

ELO, IRMA T., AND SAMUEL H. PRESTON. 1992. Effects of Early-Life Condition on Adult Mortality: A Review. *Population Index* 58(2): 186–222.

FAIG, KENNETH. 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.

FAMILY HISTORY DEPARTMENT. 1996. *The GEDCOM Standard Release 5.5.* Salt Lake City: Church of Jesus Christ of Latter-Day Saints.

FERGUSSON, DAVID M., M. E. DIMOND, JOHN L. HORWOOD, AND FREDERICK T. SHANNON. 1984. The Utilisation of Preschool Health and Education Services. *Social Science & Medicine* 11: 1173–80.

FOGEL, ROBERT W., AND DORA L. COSTA. 1997. A Theory of Technophysio Evolution, with Some Implications for Forecasting Population, Health Care Costs, and Pension Costs. *Demography* 34: 49–66.

GAVRILOV, LEONID A., AND NATALIA S. GAVRILOVA. 1991. *The Biology of Life Span: A Quantitative Approach.* New York: Harwood.

———. 1997. Parental Age at Conception and Offspring Longevity. *Reviews in Clinical Gerontology* 7: 5–12.

———. 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.

———. 1999. Season of Birth and Human Longevity. *Journal of Anti-Aging Medicine* 2: 365–66.

———. 2000. Human Longevity and Parental Age at Conception. In *Sex and Longevity: Sexuality, Gender, Reproduction, Parenthood,* ed. J.-M. Robine et al., pp. 7–31. Berlin: Springer.

———. 2001. ''The Reliability Theory of Aging and Longevity.'' *Journal of Theoretical Biology* 213: 527–45.

———. 2003a. The Quest for a General Theory of Aging and Longevity. *Science,* SAGE KE (Science of Aging and Knowledge Environment), July 16, 2003(28): 1–10.

———. 2003b. Early-Life Factors Modulating Lifespan. In *Modulating Aging and Longevity,* ed. S. I. S. Rattan, pp. 27–50. Dordrecht: Kluwer.

———. 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.

———. 2006. Reliability Theory of Aging and Longevity. In *Handbook of the Biology of Aging,* 6th ed., ed. E. J. Masoro and S. N. Austad, pp. 3–42. San Diego: Academic Press.

GAVRILOVA, NATALIA S., AND LEONID A. GAVRILOV. 1999. Data Resources for Biodemographic Studies on Familial Clustering of Human Longevity. *Demographic Research* [Online], 1(4): 1–48. www.demographic-research.org/volumes/vol1/4.

———. 2005. Search for Predictors of Exceptional Human Longevity. In *Living to 100 and Beyond,* pp. 1–49. Schaumburg, IL: Society of Actuaries.

GAVRILOVA, NATALIA S., LEONID A. GAVRILOV, GALINA N. EVDOKUSHKINA, AND VICTORIA G. SEMYONOVA. 2003. Early-Life Predictors of Human Longevity: Analysis of the 19th Century Birth Cohorts. *Annales de Demographie Historique* 2: 177–98.

GUEST, AVERY M. 1987. Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century. *Historical Methods* 20: 63–77.

HILL, MARK E., SAMUEL H. PRESTON, IRMA T. ELO, AND IRA ROSENWAIKE. 1995. Age-Linked Institutions and Age Reporting among Older African Americans. Population Research Center, University of Pennsylvania, Working Paper Series No. 95-05.

HILL, MARK E., SAMUEL H. PRESTON, AND IRA ROSENWAIKE. 2000. Age Reporting among White Americans Aged 85+: Results of a Record Linkage Study. *Demography* 37: 175–86.

HILL, MARK E., SAMUEL H. PRESTON, IRA ROSENWAIKE, AND J. F. DUNAGAN. 2000. Childhood Conditions Predicting Survival to Advanced Age among White Americans. Paper presented at the Annual Meeting of the Population Association of America, Los Angeles.

HILL, MARK E., AND IRA ROSENWAIKE. 2001. The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages. *Social Security Bulletin* 64: 45–51.

IWASHYNA, THEODORE J., JAMES X. ZHANG, DIANE S. LAUDERDALE, AND NICHOLAS A. CHRISTAKIS. 1998. A Method for Identifying Married Couples in the Medicare Claims Data: Mortality, Morbidity, and Health Care Utilization among the Elderly. *Demography* 35: 413–19.

JEUNE, BERNARD, AND JAMES VAUPEL, EDS. 1999. Validation of Exceptional Longevity. *Odense Monographs on Population Aging,* Vol. 6. Odense: Odense University Press.

KAPLAN, BERNICE A., C. G. NICHOLAS MASCIE-TAYLOR, AND JESPER BOLDSEN. 1992. Birth Order and Health Status in a British National Sample. *Journal of Biosocial Science* 24: 25–33.

KEEFE, DAVID L., KERRI MARQUARD, AND LIN LIU. 2006. The Telomere Theory of Reproductive Senescence in Women. *Current Opinion in Obstetrics and Gynecology* 18: 280–85.

KESTENBAUM, BERT. 1992. A Description of the Extreme Aged Population Based on Improved Medicare Enrollment Data. *Demography* 29: 565–80.

KESTENBAUM, BERT, AND B. RENEÉ FERGUSON. 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.

———. 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*. Schaumburg, IL: Society of Actuaries. http://library.soa.org/library-pdf/m-li05-1_XXVI.pdf.

KUH, DIANA, AND YOAV BEN-SHLOMO. 1997. *A Life Course Approach to Chronic Disease Epidemiology*. Oxford: Oxford University Press.

MAGNUS, PER, KÅRE BERG, AND TOR BJERKEDAL. 1985. The Association of Parity and Birth Weight: Testing the Sensitization Hypothesis. *Early Human Development* 12: 49–54.

MODIN, BITTE. 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: 1051–64.

NIXON, JAMES, AND JOHN PEARN. 1978. An Investigation of Socio-Demographic Factors Surrounding Childhood Drowning Accidents. *Social Science & Medicine* 12: 387–90.

PRESTON, SAMUEL H., IRMA ELO, IRA ROSENWAIKE, AND MARK HILL. 1996. African-American Mortality at Older Ages: Results of a Matching Study. *Demography* 33: 193–209.

PRESTON, SAMUEL H., AND MICHAEL R. HAINES. 1991. *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Princeton, NJ: Princeton University Press.

PRESTON, SAMUEL H., MARK HILL, AND GREG DREVENSTEDT. 1998. Childhood Conditions That Predict Survival to Advanced Ages among African Americans. *Social Science Medicine* 47: 1231–46.

ROSENWAIKE, IRA, MARK HILL, SAMUEL PRESTON, AND IRMA ELO. 1998. Linking Death Certificates to Early Census Records: The African American Matched Records Sample. *Historical Methods* 31: 65–74.

ROSENWAIKE, IRA, AND BARBARA LOGUE. 1983. Accuracy of Death Certificate Ages for the Extreme Aged. *Demography* 20: 569–85.

ROSENWAIKE, IRA, AND LESLIE F. STONE. 2003. Verification of the Ages of Supercentenarians in the United States: Results of a Matching Study. *Demography* 40: 727–39.

RUGGLES, STEVEN, MATTHEW SOBEK, TRENT ALEXANDER, CATHERINE A. FITCH, RONALD GOEKEN, PATRICIA KELLY HALL, MIRIAM KING, AND CHAD RONNANDER. 2004. *Integrated Public Use Microdata Series (IPUMS): Version 3.0*. Minneapolis: Minnesota Population Center. www.ipums.org.

SEARS, ROBERT R., ELEANOR E. MACCOBY, AND HARRY LEVIN. 1957. *Patterns of Child Rearing*. Evanston, IL: Row Peterson.

SHRESTHA, LAURA B., AND SAMUEL H. PRESTON. 1995. Consistency of Census and Vital Registration Data on Older Americans: 1970–1990. *Survey Methodology* 21: 167–77.

STATA CORP. 2005. *Stata User's Guide*. College Station, TX: Stata Press. www.stata.com.

Discussions on this paper can be submitted until July 1, 2007. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.