**The 2007 Annual Meeting of the Population Association of America**

# Problems of Mortality Measurement at Advanced Ages

**Leonid A. Gavrilov, Natalia S. Gavrilova**
Center on Aging, NORC and The University of Chicago

**Abstract**
Mortality measurement at advanced ages suffers from several problems: (1) small numbers of survivors to advanced age, which requires mixing different birth cohorts with different mortality;(2) extremely high risks of death at old ages, which make the standard assumptions of hazard rate estimates to be invalid; (3) age misreporting by old persons. The Social Security Administration Death Master File (DMF) was used to alleviate the two of the above mentioned problems and to obtain more precise monthly estimates of hazard rates after ages 85-90 years. Study of several single-year extinct birth cohorts showed that mortality grows steadily without deceleration from 80 to 102-105 years of age. Then statistical noise rapidly increases and mortality tends to decelerate. The study shows that mortality deceleration effect at advanced ages is not particularly strong when data for more homogeneous single-year birth cohorts are analyzed.

## Introduction

Accurate estimates of mortality at advanced ages are essential to improving forecasts of mortality and the population size of the oldest old age group. However, estimation of hazard rates at extremely old ages poses serious challenges to researchers:

(1) The observed mortality deceleration may be at least partially an artifact of mixing different birth cohorts with different mortality (heterogeneity effect);
(2) Standard assumptions of hazard rate estimates may be invalid when risk of death is extremely high at old ages;
(3) Ages of very old people may be highly exaggerated.

One way of obtaining estimates of mortality at extreme ages is to pool together international records of persons surviving to extreme ages with subsequent efforts of strict age validation (Robine and Vaupel 2001; Robine, Cournil et al. 2005). This approach helps to resolve the third problem mentioned above but does not allow researchers to resolve the first two problems because of inevitable data heterogeneity when data for people belonging to different birth cohorts and countries are pooled together. In this project we propose an alternative approach, which allows us to partially resolve the first two problems by compiling data for large single-year birth cohorts with hazard rates measured at narrow (monthly) age intervals. Possible ways of resolving the third problem of hazard rate estimation will also be elaborated.

The hazard rate, $\mu_x$, (or the instantaneous risk of death) is defined as follows:

$$\mu_x = -\frac{dN_x}{N_x dx}$$

where $N_x$ is a number of living individuals at age x.

One of the first empirical estimates of hazard rate was proposed by George Sacher (Sacher 1977):

$$\mu_x = \frac{1}{\Delta x}(ln\ l_{x-\frac{\Delta x}{2}} - ln\ l_{x+\frac{\Delta x}{2}}) = \frac{1}{2\Delta x}\ ln\frac{l_{x-\Delta x}}{l_{x+\Delta x}}$$

This estimate is unbiased for slow changes in hazard rate if $\Delta x \Delta \mu_x << 1$ (Sacher 1966).

A simplified version of Sacher estimate (for age intervals equal to unity) often is used in demographic and biological studies of mortality: $\mu_x = -\ln(1-q_x)$. This estimate is based on the assumption that hazard rate is constant over studied age interval, which is equal to one year for humans (see (Curtsinger, Gavrilova et al. 2005)).

At advanced ages when death rates are particularly high, the assumptions about small changes in hazard rate or a constant hazard rate within the age interval become questionable. Violation of these assumptions may lead to biased estimates of hazard rates calculated on annual basis. The narrowing of the age interval from one-year to one-month period for estimation of hazard rates is a possible way of partial resolving of the second problem mentioned above.

**Introduction**

It is now considered as an established fact that mortality at advanced ages has a tendency to deviate from the Gompertz law, so that the logistic model often is used to fit human mortality (Horiuchi, Wilmoth, 1998). The estimates of mortality force at extreme ages are difficult because of small numbers of survivors to these ages in most countries. Data for extremely long-lived individuals are scarce and subjected to age exaggeration. Traditional demographic estimates of mortality based on period data encounter well known denominator problem. More accurate estimates are obtained using the method of extinct generations (Vincent, 1951). In order to obtain good quality estimates of mortality at advanced ages researches are forced to pool data for the several calendar periods. In the Kannisto-Thatcher Database on Old Age Mortality data are aggregated for ten-year calendar periods to accumulate enough cases of survivors to older ages. Single-year life tables for many countries have very small numbers of survivors to age 100 that makes estimates of mortality at advanced ages unreliable. 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 an artifact of data heterogeneity. 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 and grows rapidly. Thus, we need more research efforts to obtain reliable estimates of mortality at advanced ages.

**Social Security Administration Death Master File as a source of mortality data for advanced ages.**

Social Security Administration Death Master File (DMF) was used in the study of mortality kinetics after ages 85-90 years. The advantage of this data source is that some birth cohorts covered by DMF could be studied by the method of extinct generations (Vincent 1951; Kannisto 1988; Kannisto 1994). Availability of month of birth and month of death information provides a unique opportunity to obtain hazard rate estimates for every month of age, which is important given extremely high mortality after age 100 years (see Table 1).

**Table 1. Variables available in the SSA Death Master File.**

| |
|---|
| 1. first, last names, SSN |
| 2. date, month, year of birth |
| 3. month, year of death |
| 4. state of the SSN issuance |
| 5. town, county, state, zip code of the last residence |
| **6.** death date verification code |

The information from the DMF was collected for individuals who lived 80 years and over and died before 2004. DMF database is unique because it represents mortality experience for one of the largest cohort of the oldest-old persons, which is readily available for survival analysis. In this study mortality measurements were made for cohorts, which are more homogeneous in respect to the period of birth and historical life course experiences.

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. 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 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.

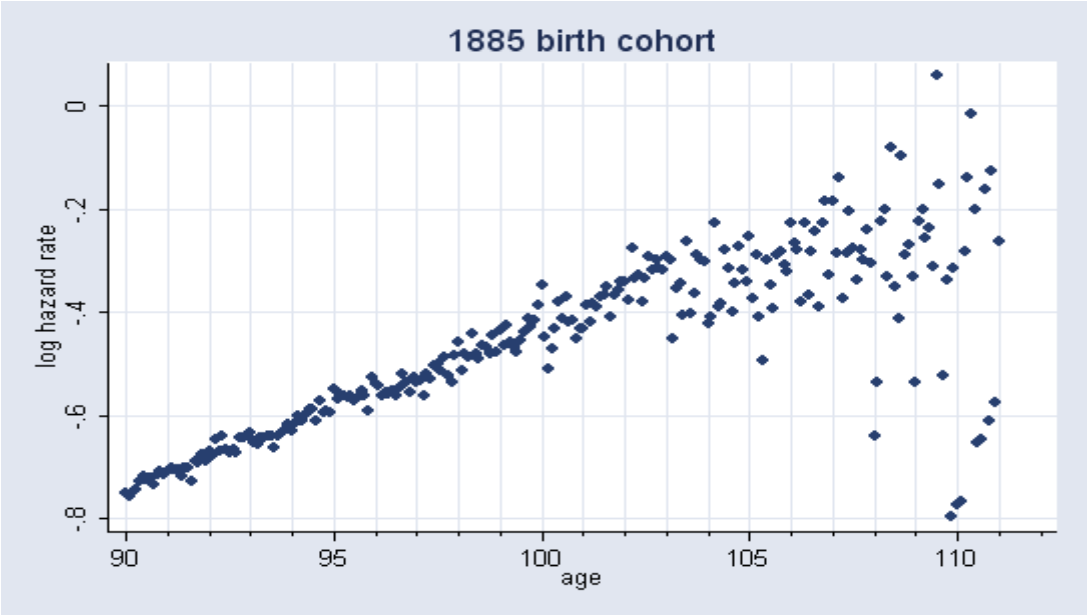


**Figure 1. Hazard rate (mortality force, year$^{-1}$) for 1885 birth cohort. Data from the Social Security Administration Death Master File. Total U.S. population.**

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 1-4.

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.

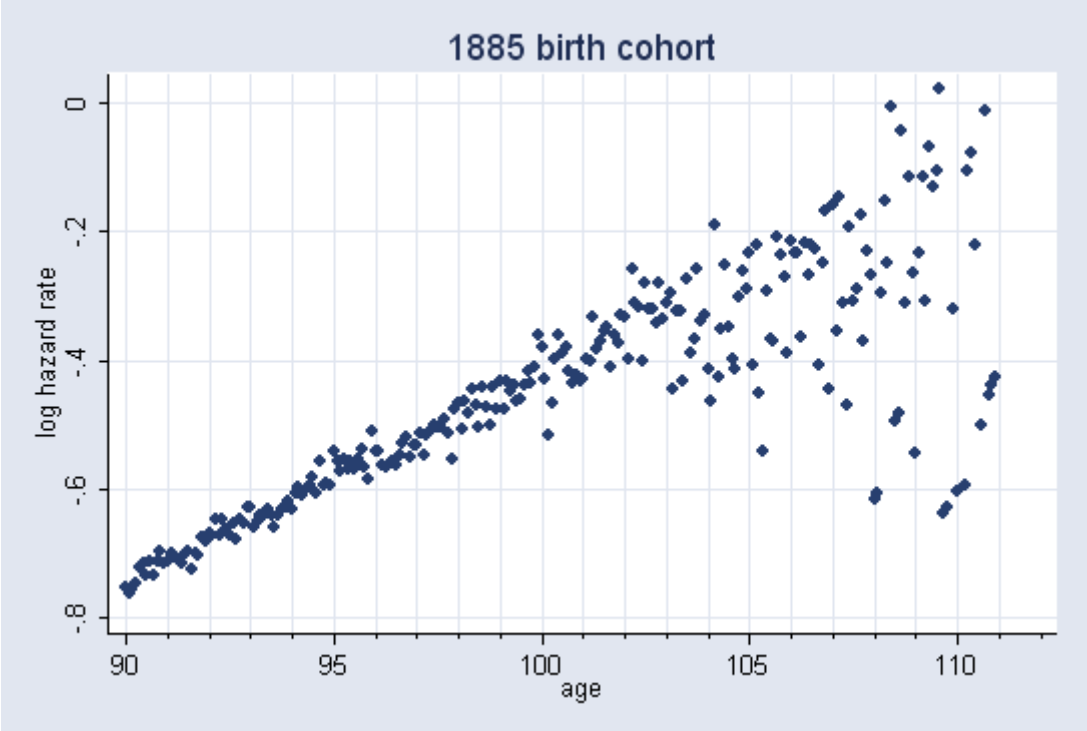


**Figure 4. Hazard rate (mortality force, year$^{-1}$) for 1885 birth cohort.**
**Data from the Social Security Administration Death Master File.**
**Less reliable data for Southern states, Puerto Rico and Hawaii are excluded.**

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 1-2).

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.

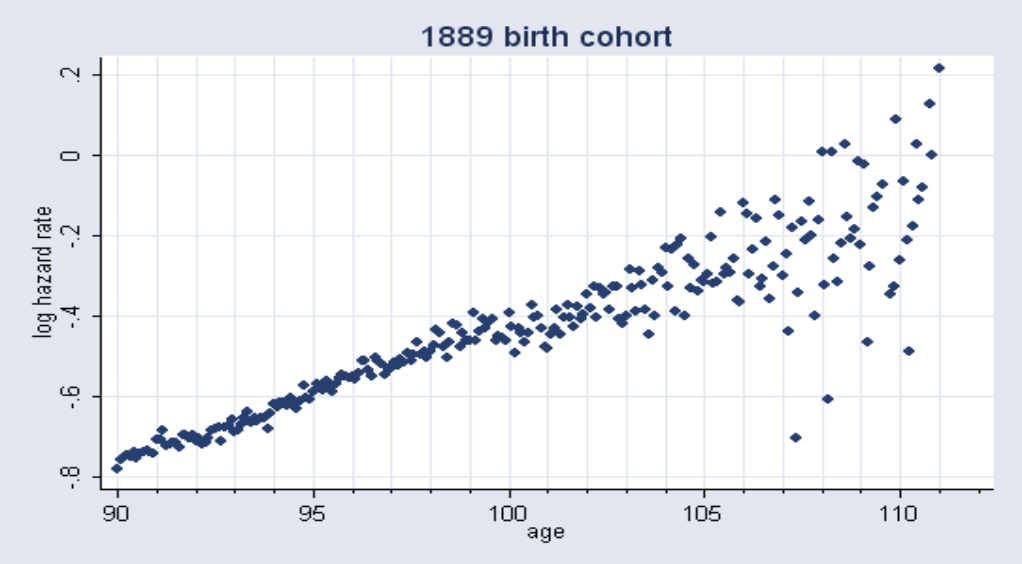


**Figure 3. Hazard rate (mortality force, year$^{-1}$) for 1889 birth cohort.**
**Less reliable data for Southern states, Puerto Rico and Hawaii are excluded.**

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.

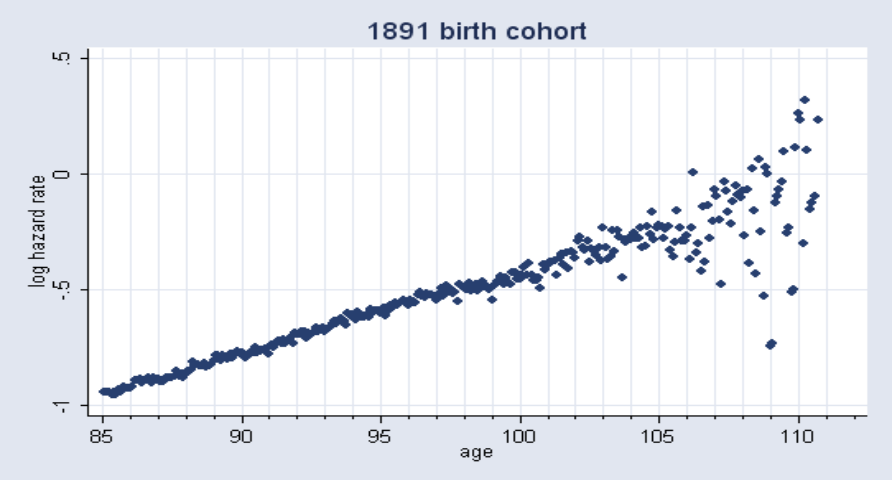


**Figure 4. Hazard rate (mortality force, year$^{-1}$) for 1891 birth cohort.**
**Less reliable data for Southern states, Puerto Rico and Hawaii are excluded.**
**Data from the Social Security Administration Death Master File.**

The result of hazard rate estimation for males and females is presented in Figure 5. 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.

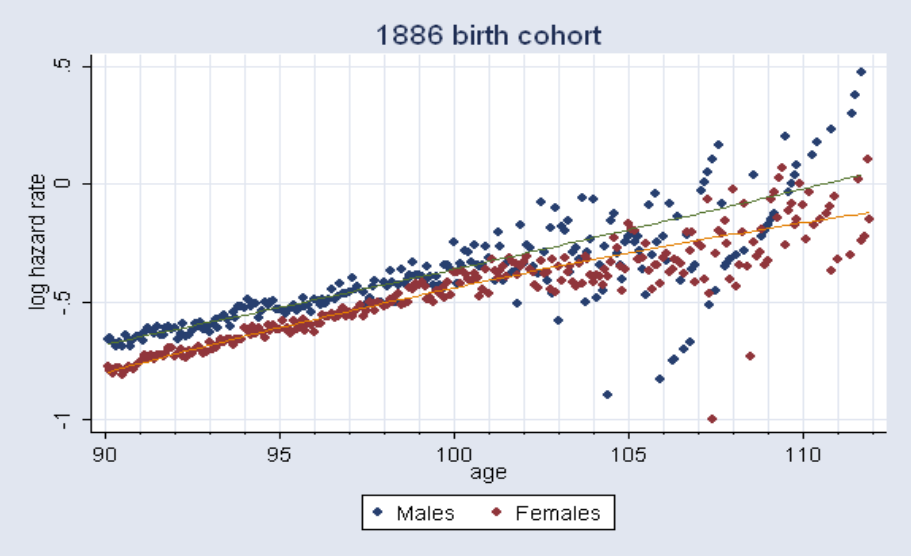


**Figure 5.**
**Hazard rate (mortality force, year$^{-1}$) for males and females from 1886 birth cohort. Data are fitted using quadratic polynomial regression. Total U.S. population.**

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 6) than estimates obtained for every month of age (Figure 5).

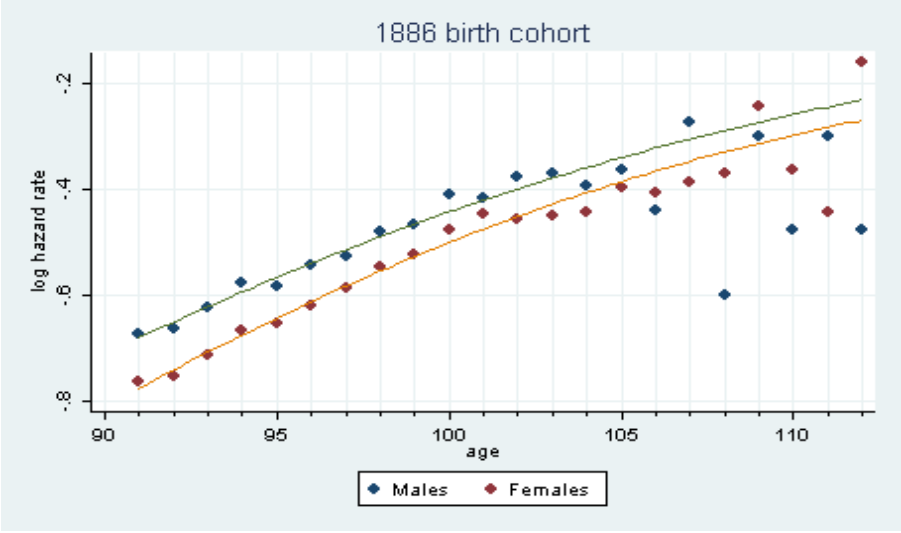


**Figure 6.**
**Hazard rate (mortality force, year$^{-1}$) for males and females from 1886 birth cohort. Lifespan is estimated in whole years. Data are fitted using quadratic polynomial regression. Total U.S. population.**

## Discussion

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. Overall these data demonstrate that mortality at advanced ages follows the Gompertz law up to the ages 102-105 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).

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.