J. theor. Biol. (2001) **213,** 547–557
doi:10.1006/jtbi.2001.2435, available online at http://www.idealibrary.com on IDEAL®

An Application of a General Branching Process in the Study of the Genetics of Aging

PETER OLOFSSON*†, OTTO SCHWALB‡, RANAJIT CHAKRABORTY§ AND MAREK KIMMEL*

*Department of Statistics, Rice University, P.O. Box 1892, Houston, TX 77251, U.S.A., ‡Equifax Decision Solutions, 1525 Windward Concourse MD42-S, Alpharetta, GA 30005, U.S.A. and §Human Genetics Center, University of Texas Health Science Center, P.O. Box 20334, Houston, TX 77225, U.S.A.

(*Received on* 10 *April* 2001, *Accepted in revised form on* 3 *August* 2001)

A general branching process model is developed to analyse familial dependence in longevity data. A general formula for the survival function of a randomly chosen sibling of an individual of a specified age is derived. The branching process model takes into account that siblings' ages may be censored. This is applied to a data set consisting of lifelengths of siblings of centenarians. Age distributions used in the branching process model are estimated from US Census data from the relevant period. It is demonstrated that there is a marked difference in the survival function according to the formula assuming no familial effect and the empirical survival function estimated from the data; thus, indicating a strong familial component.

Introduction

The aim of this paper is to explore the issue of familial clustering of longevity in families of centenarians, using life-tables and branching process modeling. This problem is a part of a more general problem of heritability of longevity, which in turn belongs to the domain of aging research. An important issue is whether there is a genetic component in aging, and it seems that the general answer is yes. Indeed, there are experimental results on various species of living organisms, indicating the existence of genetic control of aging. Similarly, there exist medical disorders of premature aging, with a proven heritable component (Arking, 1998).

However, we are seeking the answer to a very specific question about longevity in humans. We would like to find out how much of the apparent increased probability of longevity of close blood relatives of very-long-living individuals is due to sampling bias and how much is attributable to inheritance. Recently, in at least one publication (Perls *et al*., 1998) a striking apparent increase of the risk ratio (RR) of living longer was reported for siblings of centenarians from a study in Massachusetts.

Specifically, we derive the conditional probability that a sibling of a centenarian survives beyond a prescribed age, in the presence of censoring (i.e. some of the sibs are alive), assuming no familial correlation of lifespans. As an approximation, we further assume that all siblings are of the same birth cohort as the index case. Such probability calculations are important, since for any age-related phenotype such as aging, the data from random samples of a contemporary population cannot serve as a control population

† Author to whom correspondence should be addressed.

0022–5193/01/240547 + 11 $35.00/0

since it is composed of individuals of mixed birth cohorts and is subjected to censoring effects of the survival function.

We carry out our enquiry using two techniques. First, we compare the lifelength distributions of sibs of centenarians in the Massachusetts study (kindly provided by Dr Thomas T. Perls) to lifelength distributions computed based on the US Census lifetables concerning the periods when respective sibs were born. Use of the lifetables allows correcting for the cohort effect. Second, we construct a demographic model based on branching processes theory, from which we are able to compute the distributions of lifelength expected under the sampling bias stemming from the fact that some of the siblings are still alive and so their lifetimes are not known. We carry out various numerical simulations using this model, and apply it to the Massachusetts data. The general conclusion from our adjustments is that the cohort and sampling biases have a non-trivial influence, although the extent will differ in various circumstances. The model seems to be in itself an important and original application of the general (Crump–Mode–Jagers) branching process.

Description of Data

For the illustrative purpose of our computations, we use the preliminary data from a study on longevity at Beth Israel Deaconess Medical Center, Boston, MA, kindly provided by Dr Thomas T. Perls. The data consist of a list of siblings of 42 centenarians, henceforth also referred to as probands, together with the ages of these siblings, recorded at a time point in year 1996 if the subject was alive or obtained from family or other records if the subject was dead. The total number of siblings is equal to 184. In the version of the data available to us, it has not been specified as to which of the siblings were alive and which were dead. The number of siblings varies from 1 to 11 (excluding the proband). The distribution of this number is provided in Fig. 1. The distribution of ages of the siblings, excluding the proband, is provided in Fig. 2. The empirical survival function of these ages is the stepwise curve in Fig. 3.

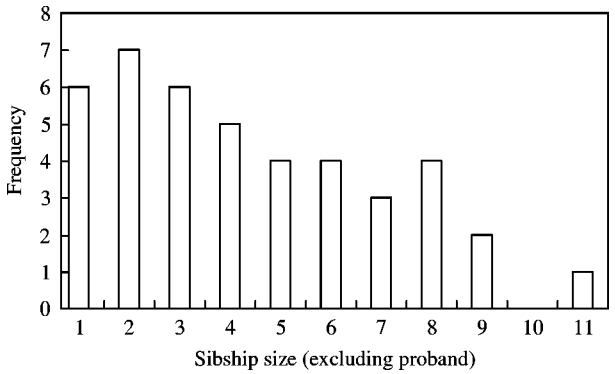

FIG. 1. Distribution of the sizes of sibships excluding the proband.

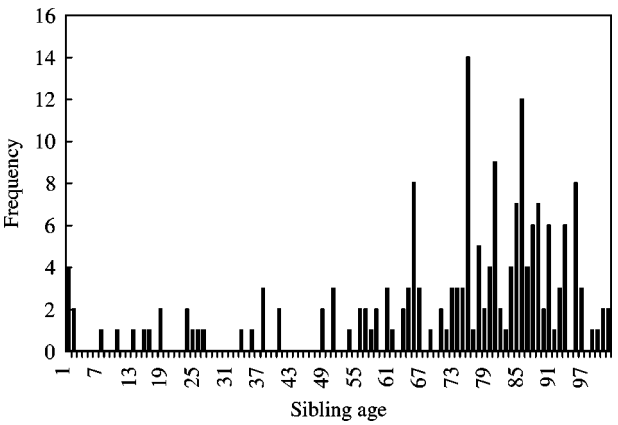

FIG. 2. Distribution of ages of the siblings, excluding the proband.

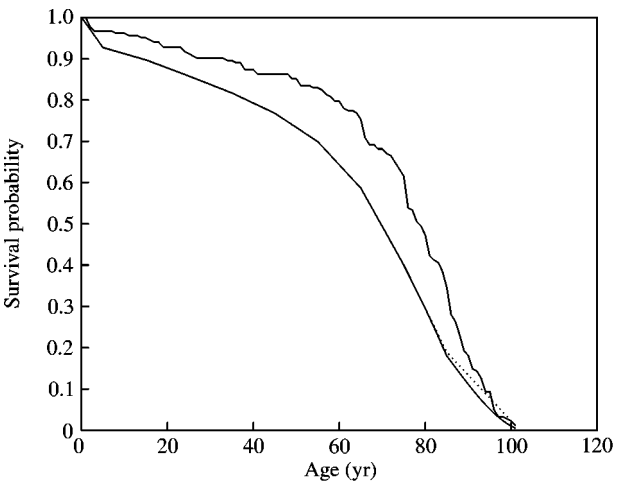

FIG. 3. Survival functions of ages of siblings of the probands, estimated from the data (~~), compared to the crude life-tables estimate (·····) and the branching process model estimate (——).

Branching Process Model of the Demographic Process

INTRODUCTION REGARDING GENERAL BRANCHING PROCESSES

This section contains a brief description of the theory of general branching processes, focusing on the needs in this particular application. This exposé follows Jagers (1992) and for more detailed descriptions, we refer the reader to this reference.

The definition of a general branching process starts from a set I of possible individuals. These are identified by descent; an individual $x \in I$ is represented as $x = (a, x_1, ..., x_n)$ with the interpretation that she is the x_n-th child of the x_{n-1}-th child... of the x_1-th child of the ancestor a. There may be one or several ancestors. With each individual $x \in I$, a reproduction process, ξ_x is associated, where $\xi_x(u)$ is the number of children up to age u. The ξ_x are assumed to be i.i.d. copies of the generic reproduction process ξ. By denoting the generic consecutive ages at childbearing by $\tau(1)$, $\tau(2), ...$ we thus get $\xi(u) = \#\{k : \tau(k) \leqslant u\}$.

Each individual x also has a birth-time, τ_x associated with it. This means that at time t, the individual x has given birth to $\xi_x(t - \tau_x)$ children. Therefore, at time t, $\sum_{x \in I} \xi_x(t - \tau_x)$ individuals have been born in the population. Note that $\xi_x(t - \tau_x) = 0$ if $t < \tau_x$, i.e. if x is not yet born at time t. This provides a method of counting the number of individuals in the population based on individual reproduction.

To be able to count or measure other aspects of the population, random characteristics are introduced. A random characteristic χ is a random process on the non-negative half-line $[0, \infty)$ with the interpretation that $\chi(u)$ is the contribution of an individual of age u (and that $\chi(u) = 0$ for $u < 0$). The χ-counted population is then defined as

$$Z_t^\chi = \sum_{x \in I} \chi_x(t - \tau_x),$$

the sum of the contributions of all individuals at time t. Note that $\chi_x(t - \tau_x) = 0$ if $t < \tau_x$, i.e. no individuals contribute before birth. By choosing $\chi(u) = I_{[0,\infty)}(u)$, an individual is counted if it is born and

$$Z_t^\chi = \sum_{x \in I} I_{\{\tau_x \leqslant t\}} = \#\{x : \tau_x \leqslant t\},$$

the number of individuals born before time t. Other choices of χ allow us to count or measure other properties. For example, if L denotes life-length, then the characteristic $\chi(u) = I_{[0,L)}(u)$ counts an individual if it is born and still alive at age u and Z_t^χ counts the number of individuals alive in the population at time t. By choosing the appropriate characteristic, a wide range of properties of the population can thus be counted or measured.

To capture the growth behavior of the population, the Malthusian parameter, α, is defined so that

$$E[\hat{\xi}(\alpha)] = 1,$$

where $\hat{\xi}$ is the Laplace transform of the reproduction process, i.e.

$$\hat{\xi}(\alpha) = \int_0^\infty e^{-\alpha t} \xi(dt).$$

In our particular application, time is naturally measured in years and the above expression becomes

$$\hat{\xi}(\alpha) = \sum_{u=0}^{\infty} e^{-\alpha u} \xi(\{u\}),$$

where $\xi(\{u\})$ is the number of children born the u-th year. Of course, in applications to human populations, this sum is always finite, $\xi(\{u\})$ being equal to 0 outside some interval $[m, M]$, the fertile years.

If the Malthusian parameter is strictly positive, it is a central result in the theory that the growth rate of the χ-counted population is exponential, $e^{\alpha t}$, and that this growth rate is the same for any characteristic χ. More precisely, the following holds:

$$e^{-\alpha t} Z_t^\chi \to E[\hat{\chi}(\alpha)] W,$$

where $E[\hat{\chi}(\alpha)]$ is the expectation of the Laplace transform of χ in the point α and W is a non-negative random variable. We do not go into the conditions of the theorem, modes of convergence or properties of W, suffice it to say here that the conditions are mild and satisfied in any conceivable application to human populations.

This convergence theorem enables us to establish results for proportions of individuals with various properties. For example, as mentioned above, the number of individuals ever born is counted with the characteristic $\chi_R(u) = I_{[0,\infty)}(u)$ and the individuals alive by $\chi_A(u) = I_{[0,L)}(u)$. At time t, the fraction of all those born who are still alive is thus

$$\frac{Z_t^{\chi_A}}{Z_t^{\chi_R}}$$

and as $t \to \infty$,

$$\frac{Z_t^{\chi_A}}{Z_t^{\chi_R}} = \frac{e^{-\alpha t} Z_t^{\chi_A}}{e^{-\alpha t} Z_t^{\chi_R}} \to \frac{E[\hat{\chi}_A(\alpha)]}{E[\hat{\chi}_R(\alpha)]}.$$

It can be easily verified that $E[\hat{\chi}_R(\alpha)] = 1$ and with F denoting the distribution function of the lifelength L, $F(t) = P(L \leqslant t)$, we obtain $E[\hat{\chi}_A(\alpha)] = 1 - \hat{F}(\alpha)$, so that the fraction of living individuals is approximately $1 - \hat{F}(\alpha)$ for large t. Similarly, the fraction of individuals with any property can be estimated by defining the appropriate characteristics for the property under consideration and also the reference population: if χ_A measures the property A and we sample from reference individuals, measured by χ_R, then the fraction of individuals with property A approaches

$$\frac{E[\hat{\chi}_A(\alpha)]}{E[\hat{\chi}_R(\alpha)]}$$

as t increases. Typically, the reference population is all those ever born or all those alive, depending on what is natural in the particular application. This leads to the concept of a stable population, loosely thought of as an old population in balanced exponential growth, where each property exists in the fraction determined by the above ratio. Hence, an individual sampled at random has probability

$$\tilde{P}(A) = \frac{E[\hat{\chi}_A(\alpha)]}{E[\hat{\chi}_R(\alpha)]} \tag{1}$$

to have property A.

It should be noted that formula (1) is based on asymptotic results and is an approximation in any application to human populations. There is also an underlying assumption of a constant growth rate which is certainly not true in general. However, our results and methods are not sensitive to small variations in growth rates and for the short period of time, we consider the assumptions are reasonable. For a comprehensive treatment of demographic projections and population evolution towards equilibrium, see Mode (1985).

ASSUMPTIONS OF THE DEMOGRAPHIC PROCESS

This work deals with an application to human populations and it is necessary to slightly alter the model to accommodate the two-sex nature of such populations, see Jagers (1982). We consider two types of individuals: females, who reproduce according to the reproduction process, ξ, and males who do not reproduce. Each new-born is female or male with equal probabilities. This means that the reproduction process, ξ, is thinned with thinning probability 1/2 to obtain the process of reproducing individuals, i.e. females, ξ_F. The thinned process has expected Laplace transform

$$E[\widehat{\xi_F}(\alpha)] = \tfrac{1}{2} E[\hat{\xi}(\alpha)],$$

so the equation defining the Malthusian parameter now becomes

$$E[\hat{\xi}(\alpha)] = 2.$$

In this application, we need to count individuals with properties that do not only depend on their own life, but also their siblings' lives. A common trick in such a case is to count mothers. For example, suppose we want to count individuals with exactly one sibling. We may then instead count mothers with exactly two children and

multiply by two. The reason for this is technical; characteristics are only allowed to depend on an individual's own life and progeny and may not depend on its ancestry. For this reason, we will only consider characteristics which are 0 for males and defined to indicate the relevant property for females. If χ_A counts females with some property A, the characteristic is

$$\chi = \chi_A I_{\{female\}},$$

which has the expected Laplace transform

$$E[\hat{\chi}(\alpha)] = \tfrac{1}{2}E[\hat{\chi}_A(\alpha)].$$

However, since the reference population consists of females only, we define the reference characteristic χ_R similarly and eqn (1) gives that the probability to have property A is

$$\tilde{P}(A) = \frac{E[\hat{\chi}_A(\alpha)]}{E[\hat{\chi}_R(\alpha)]}.$$

We assume that a female is fertile between the ages of m and M and that each year she begets a child with probability p. We assume that p does not change with age and exclude the possibility of more than one child per year (i.e. exclude twins, triplets, etc.). These assumptions are not crucial to the results but facilitate computations.

MALTHUSIAN PARAMETER

To find the Malthusian parameter, first note that the time unit is 1 yr and hence the reproduction process ξ has support on $\{m, ..., M\}$. If $X_1, X_2, ...$ are i.i.d. random variables such that $P(X_1 = 1) = 1 - P(X_1 = 0) = p$, the number of children born in year u is

$$\xi(\{u\}) = I_{[m,(L \wedge M)]}(u) X_u,$$

where $L \wedge M$ denotes the minimum of L and M. The equation defining the Malthusian parameter becomes

$$E[\hat{\xi}(\alpha)] = E\left[\sum_{u=m}^{M} \mathrm{e}^{-\alpha u}\xi(\{u\})\right]$$

$$= E\left[\sum_{u=m}^{M} \mathrm{e}^{-\alpha u} X_u I_{[m,(M \wedge L)]}(u)\right] = 2. \qquad (2)$$

In order to solve this, we need to impose a condition on the lifelength L. Let

$$g(l) = P(L = l)$$

and

$$G(M+1) = P(L \geqslant M+1) = \sum_{l \geqslant M+1} g(l)$$

to obtain

$$E[\hat{\xi}(\alpha)] = E\left[\sum_{l=m}^{M}\sum_{u=m}^{l} \mathrm{e}^{-\alpha u} X_u g(l) + G(M+1)\sum_{u=m}^{M} \mathrm{e}^{-\alpha u} X_u\right]$$

$$= p\left(\sum_{l=m}^{M}\sum_{u=m}^{l} \mathrm{e}^{-\alpha u} g(l) + G(M+1)\sum_{u=m}^{M} \mathrm{e}^{-\alpha u}\right). \qquad (3)$$

The probability p can be estimated by noting that the mean number of children of a mother is

$$\mu = E[\xi(L)]$$

$$= E\left[\sum_{l=m}^{M}\sum_{u=m}^{l} X_u g(l) + G(M+1)\sum_{u=m}^{M} X_u\right]$$

$$= p\left(\sum_{l=m}^{M} (l-m+1)g(l) + G(M+1)(M-m+1)\right).$$

THE BIASED SAMPLING FORMULA

We next turn to the problem of computing the probability that a randomly sampled sibling of a randomly sampled a-aged individual is at least b years old. We introduce the following notation:

A: age of the randomly sampled individual at the time of sampling,

S: size of the randomly sampled individual's sibship (including the individual),

R: birth rank of the randomly sampled individual

B_k: the event that the randomly sampled individual has exactly k siblings of age at least b,

E: the event that a randomly sampled sibling of the randomly sampled individual has age at least b (age $\geqslant b$) at the time of sampling.

With $\tilde{P}$ as in the previous section, we wish to compute

$$\tilde{P}(E|A=a, S\geqslant 2)=\frac{\tilde{P}(E, A=a, S\geqslant 2)}{\tilde{P}(A=a, S\geqslant 2)}. \qquad (4)$$

We consider the numerator and denominator separately. For the numerator, first consider the case $b\leqslant a$ to obtain

$$\tilde{P}(E, A=a, S\geqslant 2)=\sum_{s=2}^{M-m+1}\tilde{P}(E, A=a, S=s),$$

where

$$\tilde{P}(E, A=a, S=s)$$

$$=\sum_{r=1}^{s}\sum_{k=1}^{s-1}\tilde{P}(E, A=a, S=s, R=r, B_k),$$

where further

$$\tilde{P}(E, A=a, S=s, R=r, B_k)$$

$$=\tilde{P}(E|A=a, S=s, R=r, B_k)$$

$$\times\tilde{P}(A=a, S=s, R=r, B_k)$$

$$=\frac{k}{s-1}\tilde{P}(A=a, S=s, R=r, B_k),$$

since the probability of sampling one of the k out of $s-1$ siblings who are older than b is $k/(s-1)$. Conditioning on L now yields

$$\tilde{P}(A=a, S=s, R=r, B_k)$$

$$=\sum_{l\geqslant m+s-1}\tilde{P}(A=a, S=s, R=r, B_k, L=l).$$

We now need to define the relevant random characteristic, χ. This should count individuals with the property to be alive and of age a at sampling, to be born in a sibship of size s, to be of rank r, to have exactly k siblings older than b and to have a mother who died at age l. We do this by counting a mother at age u (dead or alive) if she at that age has a child who is alive, of age a and rank r, exactly k children who are older than b and herself died at age l. Therefore, with $\tau(r)$ denoting the age of a mother at the birth of her r-th child and λ_r the lifespan of that child, $\chi(u)$ is the indicator of the event

$$\{u=a+\tau(r),\ \lambda_r\geqslant a, L=l, S=s, B_k\},$$

which gives

$$E[\hat{\chi}(\alpha)]=\sum_{u=0}^{\infty}\mathrm{e}^{-\alpha u}E[\chi(u)]$$

$$=\sum_{u=0}^{\infty}\mathrm{e}^{-\alpha u}P(\tau(r)=u-a, \lambda_r\geqslant a, L=l, S=s, B_k),$$

where for fixed (a, s, r, l), u ranges from $m+r+a-1$ to $l+r+a-s$ if $l\leqslant M$ and from $m+r+a-1$ to $M+r+a-s$ for $l>M$.

Since λ_r is independent of everything else,

$$P(\tau(r)=u-a, \lambda_r\geqslant a, L=l, S=s, B_k)$$

$$=G(a)P(\tau(r)=u-a, L=l, S=s, B_k)$$

and

$$P(\tau(r)=u-a, L=l, S=s, B_k)$$

$$=P(B_k|\tau(r)=u-a, L=l, S=s)$$

$$\times P(\tau(r)=u-a|L=l, S=s)$$

$$\times P(S=s|L=l)P(L=l).$$

Now, given $L=l$, the number of children S has a binomial distribution with parameters $l-m+1$ and p. Hence,

$$P(S=s|L=l)=\binom{l-m+1}{s}p^s(1-p)^{l-m+1-s}.$$

Also, given $L=l$ and $S=s$, the r-th child is born at age $u-a$ if there are exactly $r-1$ children born in the age interval $[m, u-a)$ and exactly $s-r$ in the age interval $(u-a, l]$. Given $L=l$

and $S = s$, all combinations of birth ages are equiprobable and

$$P(\tau(r) = u - a | L = l, S = s) = \frac{\binom{u-a-m}{r-1}\binom{l-u+a}{s-r}}{\binom{l-m+1}{s}}$$

Both these expressions hold if $l \leqslant \mathrm{M}$, otherwise l should be replaced by M. Finally, given $\tau(r) = u - a, L = l$ and $S = s$, the number of siblings with the property to be older than b is also binomial. We have to distinguish between the cases $u - b > l$ and $u - b \leqslant l$ and obtain

$$P(B_k | \tau(r) = u - a, L = l, S = s) = \binom{s-1}{k} G(b)^k (1 - G(b))^{s-k-1} \quad \text{if} \quad u - b > l$$

and

$$P(B_k | \tau(r) = u - a, L = l, S = s) = \sum_{j=\max(0,\ s-r-l+u-b,\ k-r+1)}^{\min(s-r,\ a-b)} \binom{r-1+j}{k} \times G(b)^k (1 - G(b))^{r-1+j-k} \quad \text{if } u - b \leqslant l.$$

Again, this is for $l \leqslant M$, for $l > M$ replace l by M in the formulas. Denoting the expression for $P(B_k | \tau(r) = u - a, L = l, S = s)$ by $P(l)$ when $l \leqslant M$, $P(M)$ when $l > M$ and multiplying together yields

$$E[\hat{\chi}(\alpha)] = G(a) \sum_{s=2}^{M-m+1} \sum_{r=1}^{s} \sum_{k=0}^{s-1} \times \left(\sum_{l=m+s-1}^{M} \sum_{u=m+r+a-1}^{l+r+a-s} \mathrm{e}^{-\alpha u} p^s (1-p)^{l-m+1-s} \times \frac{k}{s-1} \binom{u-a-m}{r-1} \binom{l-u+a}{s-r} g(l) P(l) + G(M+1) P(M) \sum_{u=m+r+a-1}^{M+r+a-s} \mathrm{e}^{-\alpha u} p^s (1-p)^{l-m+1-s} \times \frac{k}{s-1} \binom{u-a-m}{r-1} \binom{M-u+a}{s-r} \right).$$

The case $b > a$ leads to a similar formula which is omitted here. Similar arguments show that the denominator is

$$\tilde{P}(A = a, S \geqslant 2) = \sum_{s} \tilde{P}(A = a, S = s) = \sum_{s,r,l} \tilde{P}(A = a, S = s, R = r, L = l).$$

Here, the relevant characteristic is the indicator of the event

$$\{u = a + \tau(r),\ \lambda_r \geqslant a, L = l, S = s\},$$

which has

$$E[\hat{\chi}(\alpha)] = G(a) \sum_{u} \mathrm{e}^{-\alpha u} P(\tau(r) = u - a, \lambda_r \geqslant a, L = l, S = s) = \sum_{u} \mathrm{e}^{-\alpha u} \binom{u-a-m}{r-1} \binom{l-u+a}{s-r} \times p^s (1-p)^{l-m+1} P(L = l),$$

so that

$$\tilde{P}(A = a, S \geqslant 2) = \sum_{u,s,r,k,l} \mathrm{e}^{-\alpha u} \binom{u-a-m}{r-1} \times \binom{l-u+a}{s-r} p^s (1-p)^{l-m+1-s} g(l).$$

Data Analysis

LIFE-TABLE ANALYSIS

The general idea of this analysis is that the distribution of age at death of an individual born around year 1896 can be obtained using the

TABLE 1
Life-table-based estimates of the distribution of the age at death of siblings of the proband

i*	M_i†	n_i‡	c_i§	q_i‖	P_i¶
1	140.2	1	0.09	0.124	0.124
2	15.4	4	0.46	0.056	0.052
3	3.3	10	0.5	0.032	0.027
4	4.45	10	0.5	0.044	0.035
5	4.9	10	0.5	0.048	0.036
6	6.2	10	0.5	0.060	0.044
7	9.4	10	0.5	0.090	0.061
8	17.45	10	0.5	0.160	0.100
9	37.85	10	0.5	0.318	0.166
10	71.56	10	0.5	0.527	0.187
11	154.2085	—	—	1	0.168

*Age interval.
†Median age-specific death rate (ASDR).
‡Length of the age interval.
§Correction factor.
‖Probability of death in interval i conditional on surviving until the beginning of this interval.
¶Probability of dying in interval i.

age-dependent death rates available from the US Census data. This distribution is estimated under the null hypothesis that the distribution of life-lengths of the siblings of the probands is identical to that in the general population. Another hypothesis is that the siblings were born around the time the probands were born. In other words, we assume that the variation of sibs' birth dates is a second-order effect.

The data at our disposal are the age-specific death rates (ASDR) taken from National Center for Health Statistics (1992). The tables published there provide death rates by age in the following age groups: [0,1), [1,5), [5,15), [15,25), [25,35), [35,45), [45,55), [55,65), [65,75), [75,85), [85, ∞). These are listed by year, for years from 1900 (the earliest available) through present. To follow the cohort of siblings, we should use data for given age-interval for the years in which the cohort reached a given age. Since ages are grouped, we take medians of rates for dates corresponding to the cohort reaching the age in a given interval. For example, for the 1896 cohort and age interval [5,15), we use the median of this age interval ASDR's for years from the interval $[1896 + 5, 1896 + 14] = [1901, 1910]$. These medians are denoted as M_i for the age interval i (Table 1).

However, our aim is to calculate probabilities P_i of death in age interval i, for a given individual of the cohort. The key to that is the conditional probability q_i (discrete hazard, cf. Cox & Oakes, 1984) of dying in age interval i, provided one survived until the beginning of this interval. Indeed, we have

$$P_i = q_i \prod_{j<i} (1 - q_j).$$

According to National Center for Health Statistics (1992, p. 5) and Newell (1988, p. 71), the probabilities q_i can be computed from rates M_i using the following relationship:

$$q_i = \frac{n_i M_i}{1 + n_i(1 - c_i)M_i},$$

where n_i is the length of age interval i and c_i is a correction factor, equal to 0.5 if the death rates are distributed uniformly over the age interval. This assumption is clearly violated for the initial intervals, and consequently, Selvin (1991, p. 313) advises the use of correction factors equal to 0.09, 0.43, 0.45, 0.47, and 0.49, for the first 5 years of life. In our case, we use $c_1 = 0.09$ and $c_2 = 0.46$, this latter being the median value for ages 1–4. The correction factors and resulting estimates of probabilities q_i and P_i are listed in Table 1. For $i = 11$ (ages 85 and above), we assume $q_{11} = 1$.

The resulting estimate of the survival function of the age at death is depicted as a dotted line in Fig. 3 (mostly coinciding with the continuous line representing the branching process estimate, see further on).

BRANCHING PROCESS ANALYSIS

The application of this model requires computing the Malthusian parameter α solving eqn (2) and then computing the probability $\tilde{P}(E_b|A = a, S \geqslant 2)$ that a randomly sampled sibling of the proband has age at least b at the time of sampling, given that the proband's age is equal to a [see expression (4)].

Table 2 depicts asymptotic characteristics of the branching process model, for several combinations of parameters. The data include the interval $[m, M]$, i.e. female fertile period, and μ, i.e. the

TABLE 2
Dependence of the Malthusian parameter and cumulative population doubling time on model parameters

m*	M†	μ‡	α§	T_d‖
15	45	5	0.032	18
15	45	3	0.014	50
20	40	5	0.031	22
20	40	3	0.014	50
15	35	5	0.038	18
15	35	3	0.017	41
20	30	5	0.037	19
20	30	3	0.016	43

*Lower bound of the fertile interval for females.
†Upper bound of the fertile interval for females.
‡Mean number of children.
§Malthusian parameter of population growth.
‖Doubling time of the cumulative number of individuals.

mean number of children in a female's lifetime. The Malthusian parameter α has been calculated using eqns (2) and (3). The doubling time is computed as $T_d = \ln(2)/\alpha$.

The values used in subsequent computations are $m = 20$, $M = 40$ and $\mu = 3$. However, the estimates of the distribution of ages of siblings are not sensitive to variation. The survival function of ages of siblings of the probands, estimated based on the branching process model is depicted by the continuous line in Fig. 3. The value of the survival function in a point t is the probability to be alive at age t, this probability being computed using the branching process formulas in conjunction with the life-table analysis from the previous section. It should be noted that the life-tables estimate and the branching process estimate coincide up to ages around 80, whereafter the branching process survival function lies entirely below the life-tables survival function. The reason for this is that censoring starts playing a role, which is accounted for in the branching process model but not in the crude life-tables based estimation. Thus, the branching process model gives stronger indications of the difference in age distributions. The magnitude of the difference changes with the age of the proband, see Fig. 6.

From Fig. 3, the relative survival probabilities can be calculated. For example, the probability to survive to age 70 is roughly 0.3 in the general population and 0.5 among the siblings of centenarians, a 70% difference. It should be kept in mind that these are rough estimates and should not be taken literally. The important observation is that, clearly, a qualitative difference exists and holds also when inherent random variation in the data set is taken into account, see next section.

FURTHER EXPLORATORY ANALYSES

Another interesting point is the shape of the survival functions for the ages of $\leqslant 5$ years. The dip of the estimated curve reflects high infant and early childhood mortality. This dip is missing from the siblings' age-based survival function, which is not unlikely to reflect underreporting of early childhood deaths in family records.

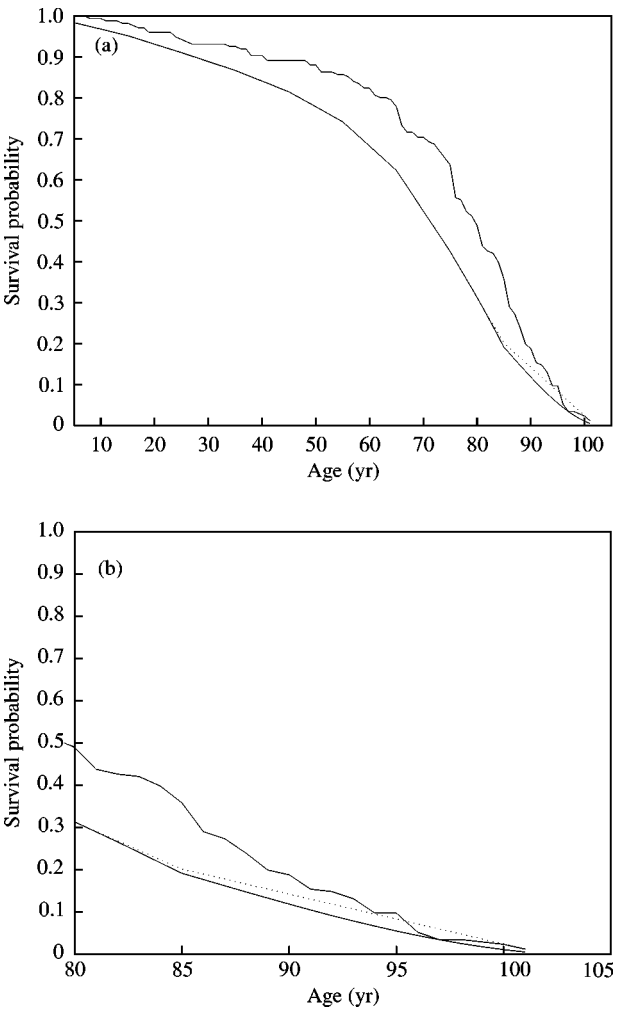

FIG. 4. Survival functions recomputed conditionally on surviving at least 5 yr, estimated from the data (~~), compared to the life-tables and branching process estimates. (a) complete plots, (b) details of the tails.

Therefore, it makes sense to recompute all survival functions conditionally on surviving at least 5 yr. The results are depicted in Fig. 4. Notice that this exercise does not alter the pattern of difference between the direct estimate and the life-tables/branching process estimates.

The empirical survival function directly obtained for sibling age data, is different from the life-tables/branching process estimate. To explore the possible significance of this difference, it seems important to determine the variability inherent in the sibling ages data. A quick method to accomplish this is by resampling. Fig. 5 depicts a family of survival functions obtained by bootstrapping from the original siblings age data, compared to the life-tables/branching process estimates. The bootstrapped family lies entirely above the branching process estimates for most of the age range, indicating that the difference observed in Figs 3 and 4 is indeed real and not due to random variation. The bootstrapped family overlaps with the estimates mostly in the lowest range of ages, but there is also some overlap in the highest range which is to be expected since there are only a few observations in the highest ages, see Fig. 2.

Finally, we explored the dependence of the branching process estimate on the age of the proband. If we assume that the proband's age was 70 and not 100, then we obtain a markedly faster decline of the estimated survival function after age 70. The reason is that a higher fraction of sibs is now still alive and generally younger. The estimated curve is depicted in Fig. 6.

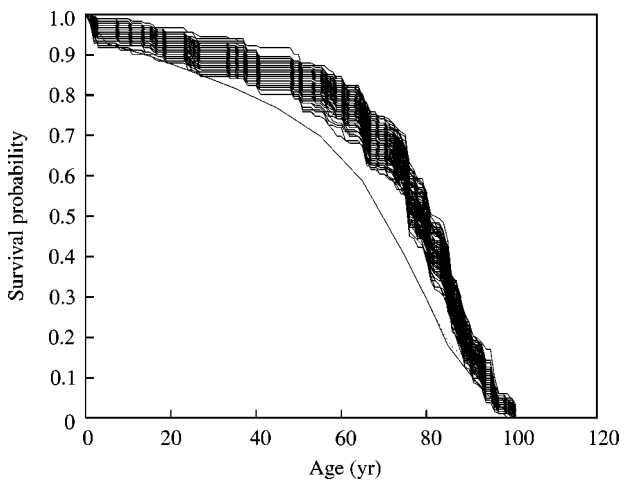

FIG. 5. A family of survival functions obtained by bootstrapping from the original siblings age data, compared to the life-tables (·····) and branching process (——) estimates.

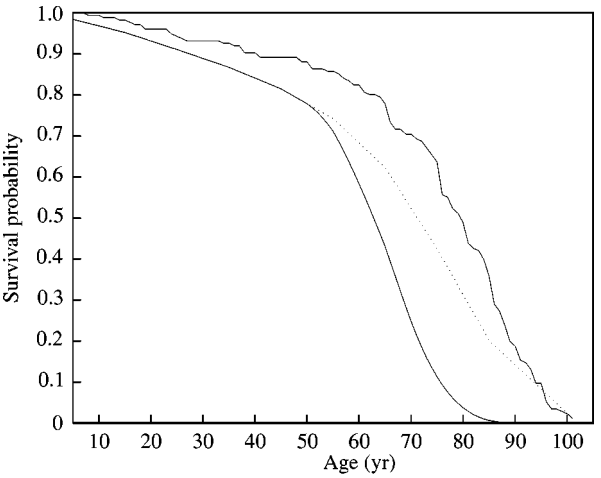

FIG. 6. The branching process estimate of the survival function of siblings' age, obtained conditionally on proband's age equal to 70 yr, compared to unchanged life-tables data-based estimates.

Discussion

The methodology developed in this paper makes it possible to compare the distribution of lifelengths of siblings of centenarians with general population data. The use of life-tables allows one to find the distribution of sibling lifelengths under the null hypothesis of no heritable component in human longevity. The observed distribution differs from the null hypothesis distribution which is consistent with the presence of a heritable component of longevity.

Further, the general branching process model allows quantitating the bias due to censoring of siblings lifelengths. For the data analysed, we demonstrate that this bias is small. This validates the conclusion attained by employing the life-tables approach.

Numerical studies indicate that the censoring bias may have a serious effect on other data sets and therefore it is worth considering. Our proposed computation is different from the one used by Perls *et al.* (1998), where the control sample consisted of lifelines of siblings of index persons

who died in 1969 at 73 years of age (i.e. with the same birth cohort of the siblings of the living centenarians at the time of the study). Our approach allows an estimation of the null distribution of the survival probabilities (i.e. without any familial correlation) directly, while Perls *et al.* (1998) estimate that empirically using a control population. In principle, risk ratios for familial aggregation of longevity, analogous to the estimates of Perls *et al.* (1998) can be computed from our Fig. 4, with bootstrap confidence intervals based on Fig. 5. However, they will not be equivalent, because of the difference in the design of generating the control sample predictions. In their case, controls are siblings of septuagenarians, while we computed the survival distribution for siblings of centenarians under the null hypothesis of no familial correlation of longevity. Our computations are of course affected by inaccuracies in the census data.

Demographic analyses usually are carried out using deterministic models, like the classical Lotka–Volterra approach. Sometimes a pseudo-stochastic interpretation, referring to expected values of an "underlying" stochastic process is provided. As argued by many, including Jagers (1991), the most convenient procedure is to actually construct this underlying process. We carry out this program in our paper. The approach has the added advantage of providing a possibility of studying extinction probabilities as well as higher moments of the process, if needed.

The calculated survival function of a randomly sampled sibling of a centenarian proband is one way of obtaining a summary statistic, which can be easily compared to data. However, the branching process approach potentially allows for a more sophisticated analysis. It seems unquestionable that there is a familial component in aging and the branching process models can be adjusted to allow for correlations between siblings' lifespans. Results for general branching processes with so-called local dependencies were developed in Olofsson (1996) and can be applied to this situation. Thus, joint distributions for lifespans in a sibship can be specified and compared to data. One practical observation is that such an analysis would probably need larger data sets to be meaningful. This is one direction for further research.

The results in this paper are descriptive in nature but it is also possible to obtain quantitative results by carrying out statistical inference. Standard non-parametric techniques for comparing two distributions can be immediately applied to our model and data, but for the type of analysis outlined above, statistical techniques would have to be examined and possibly developed. This is another direction for further research.

This work was supported by US Public Health Service research grants GM 41399 and GM 45861 (to R.C.) and GM 58545 (to P.O., R.C. and M.K.) and by the Keck's Center for Computational Biology at Rice University (MK).

REFERENCES

ARKING, R. (1998). *Biology of Aging*, 2*nd Edn*. Sunderland, MA: Sinauer.

COX, D. R. & OAKES, D. (1984). *Analysis of Survival Data*. New York: Chapman & Hall.

JAGERS, P. (1982). How probable is it to be first born? and other branching-process applications to kinship problems. *Math. Biosci.* **59,** 1–15.

JAGERS, P. (1991). The growth and stabilisation of populations. *Stat. Sci.* **6,** 269–283.

JAGERS, P. (1992). Stabilities and instabilities in population dynamics. *J. Appl. Probab.* **29,** 770–780.

MODE, C. J. (1985). *Stochastic in Demography and Their Computer Implementation*. Berlin: Springer.

National Centre for Health Statistics (1992). Vital Statistics of the United States. **2**.

NEWELL, C. (1988). *Methods and Models in Demography*. New York: Guilford Press.

OLOFSSON, P. (1996). Branching processes with local dependencies. *Ann. Appl. Probab.* **1,** 238–268.

PERLS, T. T., BUBRICK, E., WAGER, C. G., VIJG, J. & KRUGLYAK, L. (1998). Siblings of centenarians live longer. *Lancet* **351,** 1560.

SELVIN, S. (1991). *Statistical Analysis of Epidemiologic Data*, New York: Oxford University Press.