Methods Protocol for the Human Mortality Database

J.R. Wilmoth, K. Andreev, D. Jdanov, and D.A. Glei
with the assistance of
C. Boe, M. Bubenheim, D. Philipov, V. Shkolnikov, P. Vachon[1]

Table of Contents

[1] This document grew out of a series of discussions held in various locations beginning in June 2000. The four individuals listed as authors wrote the original version and/or have actively contributed to subsequent versions, including through the development of additional methods. Several others assisted with the creation of this document through their active participation in meetings and ongoing discussions via email. The authors are fully responsible for any errors or ambiguities. They thank Georg Heilmann for his assistance with the graphs.

List of Figures

List of Tables

Introduction

The Human Mortality Database (HMD) is a collaborative project sponsored by the University of California at Berkeley (United States) and the Max Planck Institute for Demographic Research (Rostock, Germany).[2] The purpose of the database is to provide researchers around the world with easy access to detailed and comparable national mortality data via the Internet.[3] When complete, the database will contain original life tables for around 35-40 countries or areas, as well as all raw data used in constructing those tables.[4] The raw data generally consist of birth and death counts from vital statistics, plus population counts from periodic censuses and/or official population estimates. Both general documentation and the individual steps followed in computing mortality rates and life tables are described here. More detailed information – for example, sources of raw data, specific adjustments to raw data, and comments about data quality – are covered separately in the documentation for each population.

We begin by describing certain general principles that are used in constructing and presenting the database. Next, we provide an overview of the steps followed in converting raw data into mortality rates and life tables. The remaining sections (including the Appendices) contain detailed descriptions of all necessary calculations.

---

[2] The contribution of UC Berkeley to this project is funded in part by a grant from the U.S. National Institute on Aging. A third team of researchers based at Rockefeller University in New York City is also working directly on this project. In addition, the project depends on the cooperation of national statistical offices and academic researchers in many countries.

[3] The HMD is accessible through either of the following addresses: www.mortality.org and www.humanmortality.de.

[4] By design, populations in the HMD are restricted to those with data (both vital statistics and census information) that cover the entire population and that are very nearly complete. We have not established precise criteria for inclusion, since we are still learning about the statistical systems of many countries. Minimally, the HMD will cover almost all of Europe, plus Australia, Canada, Japan, New Zealand, and the United States. Outside this group, there are only a few countries or areas in the world that may possess the kind of data required for the HMD (e.g., Chile, Costa Rica, Taiwan, Singapore). Nevertheless, other regions and countries are still being considered, and we do not know yet the exact list of populations that will eventually be included in the HMD. We are concerned, however, about the need to improve access to mortality information for countries that do not meet the strict data requirements of the HMD. Therefore, in addition to this project, we are also assembling a large collection of life tables constructed by other organizations or individuals. This collection is known as the Human Lifetable Database (HLD), and it will include data for many countries not covered by the HMD. The HLD is available at www.lifetable.de.

General principles

Notation and terminology for age and time

Both age and time can be either continuous or discrete variables. In discrete terms, a person "of age *x*" (or "aged *x*") has an exact age within the interval $[x, x+1)$. This concept is also known as "age last birthday." Similarly, an event that occurs "in calendar year *t*" (or more simply, "in year *t*") occurs during the time interval $[t, t+1)$. It should always be possible to distinguish between discrete and continuous notions of age or time by usage and context. For example, "the population aged *x* at time *t*" refers to all persons in the age range $[x, x+1)$ at exact time *t*, or on January 1st of calendar year *t*. Likewise, "the exposure-to-risk at age *x* in year *t*" refers to the total person-years lived in the age interval $[x, x+1)$ during calendar year *t*.

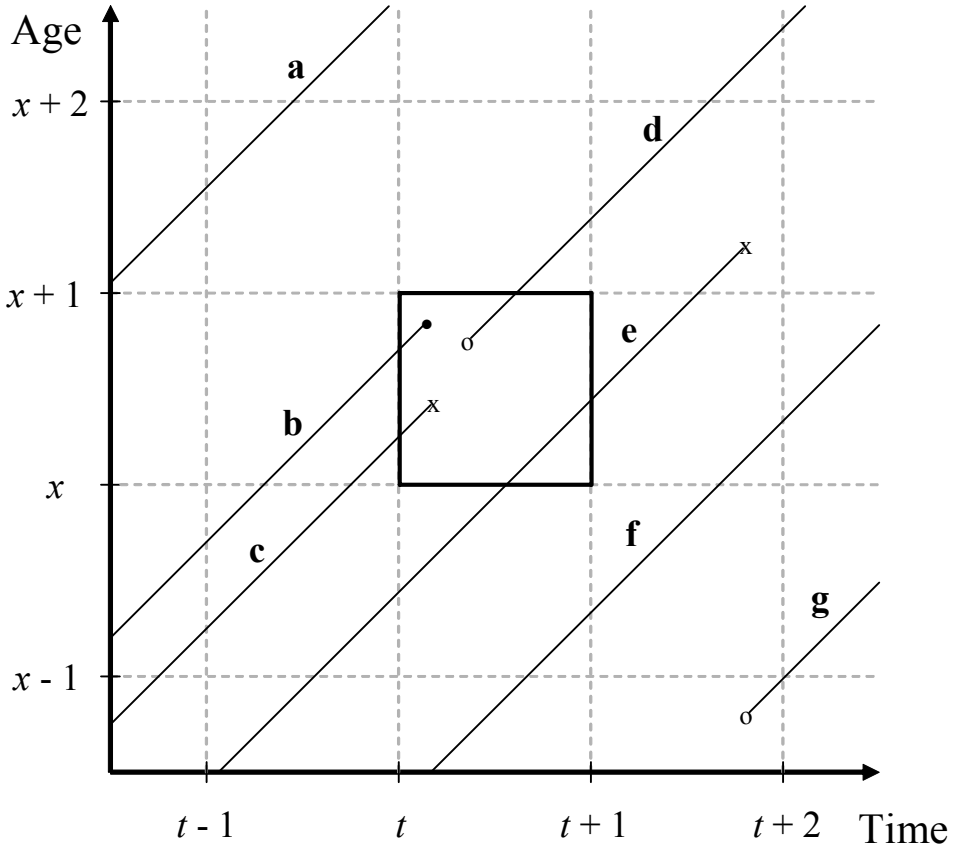
Lexis diagram

The Lexis diagram is a device for depicting the stock and flow of a population and the occurrence of demographic events over age and time. For our purposes, it is useful for describing both the format of the raw data and various computational procedures. Figure 1 shows a small section of a Lexis diagram that has been divided into 1x1 cells (i.e., one year of age by one year of time). Each 45-degree line represents an individual lifetime, which may end in death, denoted by 'x' (lines c and e), or out-migration, denoted by a solid circle (line b). An individual may also migrate into the population, denoted by an open circle (lines d and g). Other life-lines may merely pass through the section of the Lexis diagram under consideration (lines a and f).

Suppose we want to estimate the death rate for the 1×1 cell that is highlighted in Figure 1 (i.e., for age *x* to $x+1$ and time *t* to $t+1$). If the exact coordinates of all life-lines are known, then the exposure-to-risk in person-years can be calculated precisely by adding up the length of each line segment within the cell (of course, the actual length of each segment must be divided by $\sqrt{2}$, since life-lines are 45° from the age or time axes). Following this procedure, the observed death rate for this cell would be 0.91, which is the number of deaths (in this case, one) divided by the person-years of exposure (about 1.1). This is the

best estimate possible for the underlying death rate in that cell (i.e., the death rate that would be observed at that age in a very large population subject to the same historical conditions).

Figure 1. Example of a Lexis Diagram

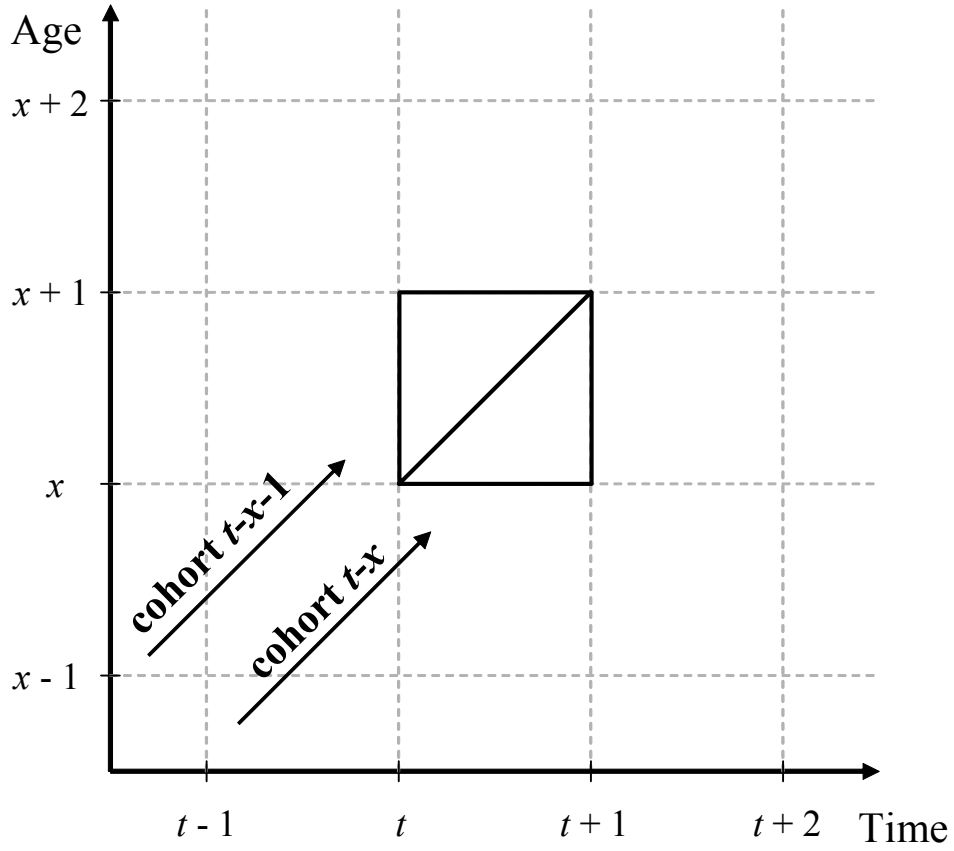


However, exact life-lines are rarely known in studies of large national populations. Instead, we often have counts of deaths over intervals of age and time, and counts or estimates of the number of individuals of a given age who are alive at specific moments of time. Considering again the highlighted cell in Figure 1, the population count at age x is 2 at time t (lines b and c) and 1 at time $t+1$ (line e). Given only this information, our best estimate of the exposure-to-risk within the cell is merely the average of these two numbers (thus, 1.5 person-years). Using this method, the observed death rate would be $1/1.5 = 0.67$, which is lower than the more precise calculation given above because the actual exposure-to-risk has been overestimated. The estimation of death rates is inevitably less precise in the absence of

information about individual life-lines, although estimates based on aggregate data using such a procedure are generally quite reliable for large populations.

Death counts are often available by age, year of death (i.e., period), and year of birth (i.e., cohort). Such counts can be represented by a Lexis triangle, as illustrated in Figure 2. Death counts at this level of detail are used in many important calculations in the HMD. One of the most important steps in computing the death rates and life tables for the HMD is to estimate death counts by Lexis triangle if these are not already available in the raw data.

Figure 2. Illustration of Lexis triangles



Standard configurations of age and time

For all data in this collection, age and time are arranged in 1-, 5-, and 10-year intervals. The configuration of a matrix of death rates (or some other quantity) is denoted by 1x1, 5x1, 5x10, etc. In this notation, the first number always refers to the age interval, and the second number refers to the time

interval. For example, 1x10 denotes a configuration with single years of age and 10-year time intervals. In the HMD, death rates and life tables are generally presented in six standard configurations: 1x1, 1x5, 1x10, 5x1, 5x5, and 5x10. Furthermore, the database includes estimates of death counts by Lexis triangle and of population size (on January 1st) by single years of age, making it possible for the sophisticated user to compute death rates and life tables in any configuration desired.[5]

All ranges of age and time describe inclusive sets of one-year intervals. For example, the age group 10-14 extends from exact age 10 up to (but not including) exact age 15, and the time period designated by 1980-84 begins at the first moment of January 1, 1980, and ends at the last moment of December 31, 1984. In addition, the following conventions are used throughout the database for organizing information by age and time:

- 5-year time intervals begin with years ending in '0' or '5' and finish with years ending in '4' or '9';
- 10-year time intervals begin with years ending in '0' and finish with years ending in '9';
- incomplete 5- or 10-year time intervals are included in presentations of death rates or life tables if data are available for at least 2 years (at either the beginning or the end of the series);
- for raw data, data in one-year age groups are always provided up to the highest age available (followed by an open age interval only if more detailed data are not available);
- for all data on country pages, one-year age groups stop at age 109, with a final category for ages 110 and above;
- for 5-year age groups, the first year of life (age 0) is always separated from the rest of its age group (ages 1-4), and the last age category is for ages 110 and above. Thus, a 5x1 configuration contains data for single years of time with (typically) the following age intervals: 0, 1-4, 5-9, 10-14, …, 105-109, 110+.

[5] In future versions of this database, we hope to add an interactive component that would allow a user to request death rates or life tables in a wider variety of age-time configurations.

It is important to note that the data shown on country pages by single years of age up to 110+ are sometimes the product of aggregate data (e.g., five-year age groups, open age intervals), which are split into finer age categories using the methods described here. Although there are some obvious advantages to maintaining a uniform format in the presentation of death rates and life tables, it is important not to interpret fictitious data literally. In all cases, the user must take responsibility for understanding the sources and limitations of all data provided here.

Female / male / total

In this database, life tables and all data used in their construction are available for women and men separately and together. In most cases, a single file contains columns labeled "female," "male," and "total" (note that this is alphabetical order). However, in the case of life tables, which already contain several columns of data for each group, data for these three groups are stored in separate files.

Raw data for women and men are always pooled prior to making "total" calculations. In other words, death rates and other quantities are not merely the average of the separate values for females and males. For this reason, all "total" values are affected by the relative size of the two sexes at a given age and time.

Periods and cohorts

Raw data are usually obtained in a period format (i.e., by the year of occurrence rather than by year of birth). Deaths are sometimes reported by age and year of birth, but the statistics are typically collected, published, and tabulated by year of occurrence. Although raw data are presented here in a period format only, death rates and life tables are provided in both formats if the observation period is sufficiently long to justify such a presentation. Death rates are given in a cohort format (i.e., by year of birth) if there are at least 30 consecutive calendar years of data for that cohort. Cohort life tables are

presented if there is at least one cohort observed from birth until extinction.[6] In that case, life tables are provided for all extinct cohorts and for some almost-extinct cohorts as well.[7]

Adjustments to raw data

Most raw data are not totally "clean" and require various adjustments before being used as inputs to the calculations described here. The most common adjustment is to distribute persons of unknown age (in either death or census counts) across the age range in proportion to the number of observed individuals in each age group. Another common adjustment is to split aggregate data into finer age categories – in the case of death counts, from 5x1 to 1x1 data, and from 1x1 data to Lexis triangles. These two common procedures are described later in this document.

Format of data files

Raw data for this database have been assembled from various sources. However, all raw data have been assembled into files conforming to a standardized format. There are different formats for births, deaths, census counts, and population estimates. The raw data files on the web page are always presented in one of these standardized formats. Output data – such as exposure estimates, death rates, and life tables – are also presented in standardized formats.

Steps for computing mortality rates and life tables

There are six steps involved in computing mortality rates and life tables for the core section of the HMD. Computational details are provided in later sections of this document, including the appendices. Here is just an overview of the process:

1. Births. Annual counts of live births by sex are collected for each population over the longest possible time period. At a minimum, a complete series of birth counts is needed for the time period over which mortality rates and period life tables are computed. These counts are used mainly for

[6] An extinct cohort is one whose members are assumed to have all died by the end of the observation period. A rule for identifying the most recent extinct cohort is given later.
[7] A simple decision rule is used to determine when it is acceptable to compute life tables for almost-extinct cohorts. In such cases, death rates for ages not yet observed are based on the average experience of previous cohorts. A detailed description of these procedures is given in a later section.

estimating the size (on January 1st of each year) of individual cohorts from birth until the time of their first census, and for other adjustments based on relative cohort size.

2. Deaths. Death counts are collected at the finest level of detail available – ideally, by individual triangles of the Lexis diagram. Sometimes, however, death counts are available only for 1x1 Lexis squares or 5x1 Lexis rectangles. Before making subsequent calculations, deaths of unknown age are distributed proportionately across the age range, and aggregated deaths are split into finer age categories. Additional adjustments or *ad hoc* estimations may be necessary, depending on the characteristics of the raw data for a particular population (any such adjustments are described in the documentation for that population).
3. Population size. Below age 80, estimates of population size on January 1st of each year are either obtained from another source (most commonly, official estimates) or derived using intercensal survival. In most cases, all available census counts are collected for the time period over which mortality rates and life tables are computed. The maximum level of age detail is always retained in the raw data and used in subsequent calculations. When necessary, persons of unknown age are distributed proportionately into other age groups before making subsequent calculations. Above age 80, population estimates are derived by the method of extinct generations for all cohorts that are extinct (see below for extinction rule), and by the survivor ratio method for non-extinct cohorts who are older than age 90 at the end of the observation period. For non-extinct cohorts aged 80 to 90 at the end of the observation period, population estimates are obtained either from another source or by applying the method of intercensal survival.
4. Exposure-to-risk. Estimates of the population exposed to the risk of death during some age-time interval are based on annual (January 1st) population estimates, with a small correction that reflects the timing of deaths during the interval.
5. Death rates. For both periods and cohorts, death rates are simply the ratio of death counts and exposure-to-risk estimates in matched intervals of age and time.

6. Life tables. Period death rates are converted to probabilities of death by a standard method. Cohort probabilities of death are computed directly from raw data, but they are related to cohort death rates in a consistent way. These probabilities of death are used to construct life tables.

Common adjustments to raw data

In this section, we give formulas for four common adjustments to raw data: 1) redistributing deaths of unknown age, 2) splitting 1x1 deaths counts into Lexis triangles, 3) splitting 5x1 deaths counts into 1x1 data, and 4) splitting deaths counts in open age intervals into 1x1 data.[8]

Distributing deaths of unknown age

The most common adjustment to raw data involves distributing observations (either deaths or census counts) where age is unknown into specific age categories. In general, such observations are distributed proportionally across the age range.

For example, suppose that death counts are available for individual triangles of the Lexis diagram but that age is unknown for some number of deaths. Formally, let

$D_L(x,t)$ = number of lower-triangle deaths recorded among those aged $[x, x+1)$ in year t ;

$D_U(x,t)$ = number of upper-triangle deaths recorded among those age $[x, x+1)$ in year t ;

$D_{UNK}(t)$ = number of deaths of unknown age in year t ;

and $D_{TOT}(t)$ = total number of deaths in year t

$$= \sum_x \left[D_L(x,t) + D_U(x,t)\right] + D_{UNK}(t) \ .$$

[8] In recent years, some national statistical offices have begun reporting deaths by year of occurrence as well as by year of registration (which may differ if registration was delayed). In such cases, we tabulate deaths according to the year in which they occurred. If data are not available by Lexis triangle, we split them into triangles using the methods described in this document. If deaths that were registered late are available in the same format as other deaths, we sum the two sets of data first and then split them into triangles.

Then, the following pair of equations redistributes deaths of unknown age proportionally across upper and lower Lexis triangles over the full age range:

$$D_L^*(x,t) = D_L(x,t) + D_{UNK}(t)\frac{D_L(x,t)}{\sum_x \left[D_L(x,t) + D_U(x,t)\right]} = D_L(x,t)\cdot\left(\frac{D_{TOT}(t)}{D_{TOT}(t) - D_{UNK}(t)}\right) \qquad (1)$$

and $$D_U^*(x,t) = D_U(x,t) + D_{UNK}(t)\frac{D_U(x,t)}{\sum_x \left[D_L(x,t) + D_U(x,t)\right]} = D_U(x,t)\cdot\left(\frac{D_{TOT}(t)}{D_{TOT}(t) - D_{UNK}(t)}\right) \qquad (2)$$

for all ages x in year t.[9]

Obviously, these calculations typically result in non-integer death "counts" for individual ages and Lexis triangles. In fact, such numbers are no longer true counts but rather estimated counts. However, since they are our best estimates of actual death counts, it is appropriate to use them in all subsequent calculations. In all formulas given below, it is assumed that deaths of unknown age have been distributed proportionally, if needed, and the superscript * used in this section is suppressed for sake of simplicity.

When raw death counts are available in a 1x1 or 5x1 format, deaths of unknown age (if any) are distributed across the existing age groups before splitting the raw counts into Lexis triangles, as described below. Note, however, that the final result of these calculations would not change if aggregate data were first split into finer age categories before redistributing deaths of unknown age. In other words, the ordering of these procedures does not matter.

Like death counts, census tabulations may contain persons of unknown age. If needed, a similar adjustment is made before proceeding with the calculations used for estimating population on January 1st as described in a later section.

[9] This adjustment reflects an assumption that the probability of age not being reported is independent of age itself.

Splitting 1x1 death counts into Lexis triangles

Death counts are often available only in 1x1 Lexis squares and not in Lexis triangles.[10] Since many of our subsequent calculations are based on Lexis triangles, it is necessary to devise a method for splitting 1x1 death data into triangles, when necessary. In general, the proportion of deaths in lower and upper Lexis triangles varies with age, as shown by the regression model presented later in this section (see also Vallin, 1973). Nevertheless, an adequate procedure in many cases is simply to assign half of each 1x1 death count to the corresponding lower and upper triangles, since errors of overestimation for one triangle (in a lower-upper pair) are typically balanced by errors of underestimation for the other triangle in almost all subsequent calculations. This simple procedure was applied successfully to the analysis of mortality above age 80 in the Kannisto-Thatcher database (Andreev, 2001).

However, for a collection of mortality data in both period and cohort formats covering the entire age range, a more complicated procedure is needed for at least two reasons: (1) deaths in the first year of life are heavily concentrated in the lower triangle and should not be split in half, and (2) at any age, the distribution of deaths across the two triangles is affected by the relative size of the two cohorts (and sometimes by historical events as well). The second point is especially important in situations where there are rapid changes in cohort size due to marked discontinuities in the birth series, as occurs in times of rapid social change (for example, at the beginning or end of a major war). Once the procedure for splitting 1x1 deaths is modified to take these matters into account, it is only a small step further toward a complete model that adjusts for several factors that are known to affect the distribution of deaths by Lexis triangle.

For these reasons, we have developed a regression equation for use in splitting 1x1 deaths into Lexis triangles. The equation is based on a multiple regression analysis of data for three countries, which

[10] Sometimes death counts are available only by period-cohort parallelogram (i.e., holding calendar year and birth cohort constant, but covering more than one age year). Within each single year birth cohort, these deaths are simply split in half into the two respective Lexis triangles. Similarly, deaths counts may be available by age-cohort parallelogram (i.e., age and birth cohort are constant, but the parallelogram covers more than one calendar year), in which case we also split the deaths in half into Lexis triangles (for year t and year t+1).

is described more fully in Appendix A. The equation is expressed in terms of the proportion of deaths that occur in the lower triangle. In general, we denote this proportion as follows:

$$\pi_d(x,t) = \frac{D_L(x,t)}{D_L(x,t) + D_U(x,t)} \quad . \qquad (3)$$

When the values of $D_L(x,t)$ and $D_U(x,t)$ are not known, our task is to derive an estimated proportion in the lower triangle, denoted $\hat{\pi}_d(x,t)$. From this quantity, we compute estimates of lower- and upper-triangle deaths: $\hat{D}_L(x,t) = \hat{\pi}_d(x,t) \cdot D(x,t)$ and $\hat{D}_U(x,t) = D(x,t) - \hat{D}_L(x,t) = \left[1 - \hat{\pi}_d(x,t)\right] \cdot D(x,t)$, where $D(x,t)$ is the observed number of deaths in the 1x1 Lexis square.

The equation for estimating $\pi_d(x,t)$ differs by sex. For women, the equation is as follows:

$$\begin{aligned} \hat{\pi}_d(x,t) = {} & 0.4710 + \hat{\alpha}_x^F + 0.7372 \cdot \left[\pi_b(x,t) - 0.5\right] \\ & + 0.1025 \cdot I(t = 1918) - 0.0237 \cdot I(t = 1919) \\ & - 0.0112 \cdot \log IMR(t) \\ & - 0.0688 \cdot \log IMR(t) \cdot I(x = 0) \\ & + 0.0268 \cdot \log IMR(t) \cdot I(x = 1) \\ & + 0.1526 \cdot \left[\log IMR(t) - \log 0.01\right] \cdot I(x = 0) \cdot I(IMR < 0.01) \end{aligned} \quad . \qquad (4)$$

In this equation, "log" refers to the natural logarithm. The indicator function, $I(.)$, equals one if the logical statement within parentheses is true and zero if it is false. Dummy variables for years 1918 and 1919 are included to reflect the strong impact of the worldwide Spanish flu epidemic on the distribution of deaths within those two years. The estimated age effects, $\hat{\alpha}_x^F$, for the female version of the equation, are given in Table A-1a (in Appendix A) under the column for Model VI. Except for ages 0 and 1, the same age coefficient is used for more than one single-year age within a broader age group, and the coefficient for the age group 100-104 is used for all ages above 100 years. The birth proportion, $\pi_b(x,t)$,

is defined formally as follows:[11]

$$\pi_b(x,t) = \frac{B(t-x)}{B(t-x) + B(t-x-1)} \ , \qquad (5)$$

where $B(t)$ is the number of births (sexes combined) occurring in year *t* in the same population.[12] Wherever the available birth series is incomplete, we set $\pi_b(x,t) = 0.5$.

The infant mortality rate (sexes combined) is found using a method proposed by Pressat (1980):

$$IMR(t) = \frac{D(0,t)}{\frac{1}{3}B(t-1) + \frac{2}{3}B(t)} \ . \qquad (6)$$

Note that the infant mortality rate can be computed in this manner before splitting 1x1 deaths into triangles. If $B(t)$ and $D(0,t)$ are known but $B(t-1)$ is unknown, then we set $B(t-1) = B(t)$ to calculate $IMR(t)$. In general, the historical decline in infant mortality has been associated with a higher proportion of deaths in the lower triangle (relative to the upper triangle) across the age range, except at age 1. At age 0, the decline in infant mortality is associated with a rapidly increasing concentration of deaths within the lower triangle, until the IMR falls below one percent. Below that level, the historical trend reverses itself, and the proportion of infant deaths in the lower Lexis triangle tends to fall.

For men, the equation for estimating $\pi_d(x,t)$ is as follows:

$$\begin{aligned}\hat{\pi}_d(x,t) = {} & 0.4838 + \hat{\alpha}_x^M + 0.6992 \cdot \left[\pi_b(x,t) - 0.5\right] \\ & + 0.0728 \cdot I(t=1918) - 0.0352 \cdot I(t=1919) \\ & - 0.0088 \cdot \log IMR(t) \\ & - 0.0745 \cdot \log IMR(t) \cdot I(x=0) \\ & + 0.0259 \cdot \log IMR(t) \cdot I(x=1) \\ & + 0.1673 \cdot \left[\log IMR(t) - \log 0.01\right] \cdot I(x=0) \cdot I(IMR < 0.01)\end{aligned} \ . \qquad (7)$$

[11] The birth proportion provides information about the relative size of two successive birth cohorts, who both pass through the age interval [*x*, *x*+1) during calendar year *t*. More precisely, it expresses the original size of the younger cohort (passing through the lower triangle of a 1x1 Lexis square) as a proportion of the total births for the two cohorts. Although this number measures the relative size of the two cohorts at birth, it can also serve as a useful indicator of their relative sizes at later ages.

[12] In the case of a country or area that has undergone territorial changes, it is important to adjust the birth series so that it refers always to the same population. See Appendix D for a general discussion of how we deal with changes in population coverage.

In this equation, $\pi_b(x,t)$ and $IMR(t)$ are the same as in the female equation, since each is based on the total population. However, the age coefficients (as well as all other coefficients) are different and are given in Table A-1b (in Appendix A) under Model VI.

Splitting 5x1 death counts into 1x1 data

Death counts in a 5x1 configuration are split into 1x1 data using cubic splines fitted to the cumulative distribution of deaths within each calendar year. In principle, the same or a similar method could be applied to any configuration of death counts by age.[13] The method used here requires only that the raw data include death counts for the first year of life and for the first five years of life. Other than these two restrictions, it does not matter whether the raw data are strictly in five-year age groups (after age five) or in some other configuration. Also, there can be an open age interval above 90, 100, or some other age. The spline method is used to split death counts for all ages below the open age interval. Details of the computational methods are given in Appendix B.

Splitting death counts in open age intervals into Lexis triangles

In some cases the raw data provide no age detail on death counts above a certain age *x*. Instead, we know only the total number of deaths in this open age interval for some calendar year *t*, which we denote ${}_{\infty}D_x(t)$. In these situations we need a method for splitting ${}_{\infty}D_x(t)$ into finer age categories. One possibility would be to split death counts in the open age interval into 1x1 data and then to apply the method described earlier for splitting 1x1 death counts into Lexis triangles. However, the method for splitting the open age interval itself is inevitably arbitrary and imprecise, and it seems that little would be gained by such a 2-step procedure. Therefore, our method splits ${}_{\infty}D_x(t)$ immediately into Lexis triangles.

[13] For some populations, we have death counts by period-cohort parallelograms covering five cohorts (e.g., deaths in year *t* for the *t*-9 to *t*-5 birth cohorts who will complete ages 5-9 in year *t*). In this case, we use the cubic spline method described here to split these deaths into single birth cohorts (see Appendix B for more details).

In order to distribute deaths in the open age interval, we fit the Kannisto model of old-age mortality (Thatcher et al., 1998) to death counts for ages $x^* - 20$ and above, where x^* is the lower boundary of the open age group (e.g., 80, 90, 100), thus treating death counts within a period as though they pertain to a closed cohort. We then use the fitted model to extrapolate death rates by Lexis triangle within the open age interval and use those rates to derive the number of survivors at each age. For details, see Appendix C.

Population estimates (January 1st)

We describe four methods for deriving age-specific estimates of population size on January 1st of each year: 1) linear interpolation, 2) intercensal survival, 3) extinct cohorts, and 4) survivor ratios. For most of the age range, we use either linear interpolation of population estimates from other sources[14] or intercensal survival methods. At ages 80 and older, we use population estimates computed using the methods of extinct cohorts and survivor ratios (except for those cohorts who are younger than age 90 at the end of the observation period). We describe the four methods separately. In case of territorial changes (or other changes in population coverage) during the time period covered by HMD, adjustments to these methods are described in Appendix D.

Linear interpolation

In some cases, the available population estimates from other sources are for some date other than January 1st (e.g., mid-year estimates). When the period between one population estimate and the next (or a population estimate and a census count) is one year or less, we use linear interpolation to derive the January 1st population estimate.[15] When the period between population counts is greater than one year (e.g., census counts), we employ intercensal survival.

[14] The main criteria for using population estimates from another source are that they are available and that they are believed to be reliable.

[15] We calculate the population as of January 1st of year *t* as a weighted average of the estimates in years *t* and *t*-1, where the weights are based on the proportion of the year between January 1st and the date of the available estimate. For example, if we have October 1st estimates, then the January 1st population at age *x* is calculated as:

$$P(x,01.01.YYYY) = 0.75 \cdot P(x,01.10.YYYY\text{-}1) + 0.25 \cdot P(x,01.10.YYYY) \quad .$$

Intercensal survival methods

Intercensal survival methods provide a convenient and reliable means of estimating the population by age on January 1st of every year during the intercensal period. There are two cases: (1) pre-existing cohorts (i.e., those already alive at the time of the first census), and (2) new cohorts (i.e., those born during the intercensal interval). We develop formulas for these two situations separately by first considering the simple case of a country that conducts censuses every five years on January 1st. We then propose a more general method that can be used for censuses occurring at any time of the year and for intercensal intervals of any length.

Specific example

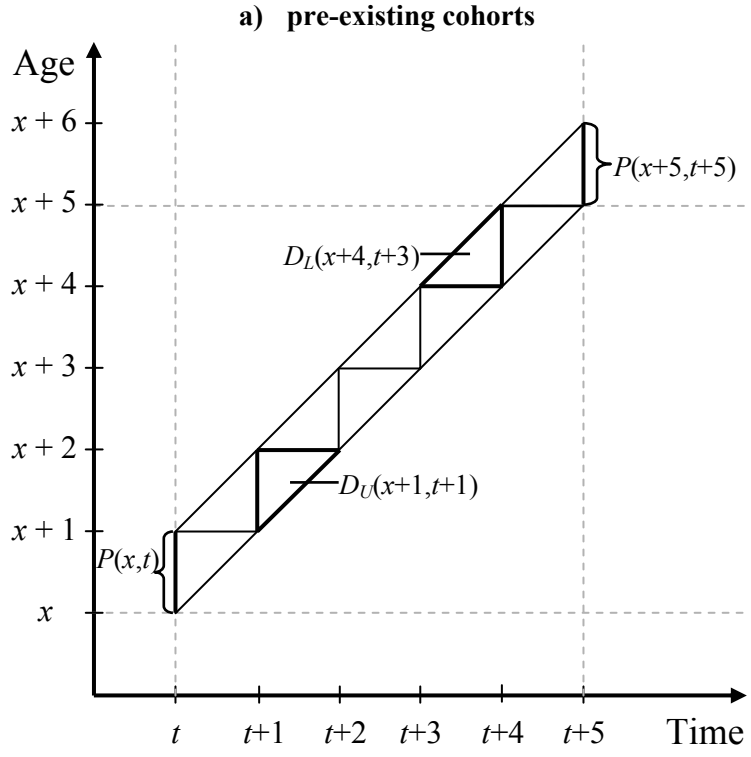
Suppose that a country conducts censuses every five years, and suppose that each census occurs on January 1st. Therefore, population estimates by single years of age are available at five-year intervals, but no comparable estimates are available for intervening years.

1. Pre-existing cohorts

The Lexis diagram in Figure 3a depicts a cohort who is already alive at the time of the first census. The cohort aged x at time t is followed through time for 5 years. Suppose that all deaths in the population are recorded with a relatively high level of detail, such that for each year in the intercensal period, death counts are available by both age and year of birth. Thus, it is known with some precision how many life-lines ended by death in each of the small triangles shown in this figure.

At the beginning or end of the data series, we cannot use linear interpolation because there are not two data points (e.g., the last population estimate in the series is for July 1st of year t). In these cases, we use "pre-censal" or "post-censal" estimation (see p. 25) to derive the January 1st estimate (i.e., by adding or subtracting deaths for each cohort).

Figure 3. Intercensal survival method (example)



The information represented by Figure 3a can be used to estimate the size of the cohort on January 1st of each intercensal year. The simplest procedure consists merely of subtracting death counts from the initial census count to obtain cohort population estimates on January 1st of each succeeding year. Unfortunately, the final step of such a computation usually yields an estimate of cohort size at time t+5 that differs from the number given by the corresponding census. This inconsistency is caused by two factors: migration and error. Although both of these factors tend to be small relative to cohort size (at least for national populations), as a matter of principle they should not be ignored. The standard method consists of distributing implied migration/error uniformly over the parallelogram shown in Figure 3a. Then, estimates of cohort size for intercensal years are found by subtracting, from the initial census count, both the observed death counts and an estimate of net migration/error.

Formally, the procedure can be described as follows. Let $C_1(x)$ equal the census count for

persons aged $[x, x+1)$ on January 1st of year *t*. Assuming that there is no migration or error, note that

$$C_1(x) = \sum_{i=0}^{\infty} \left[D_U(x+i, t+i) + D_L(x+i+1, t+i) \right] . \quad (8)$$

This formula resembles one that is used for estimating population sizes at older ages (the extinct cohort method, see below).

Using census information about the size of a cohort at time *t*, we can *estimate* its size at the time of the next census, $t+5$, by the following formula:

$$\hat{C}_2(x+5) = C_1(x) - \sum_{i=0}^{4} \left[D_U(x+i, t+i) + D_L(x+i+1, t+i) \right] . \quad (9)$$

However, if there is any migration into or out of this cohort during the intercensal period, or any error in the recording of census or death counts, this estimate will differ from the actual count at the time of the next census, $C_2(x+5)$. By definition, total migration/error is equal to the observed cohort size at the second census minus its estimated size, $\hat{C}_2(x+5)$. We call this difference Δ_x:

$$\Delta_x = C_2(x+5) - \hat{C}_2(x+5) \ . \quad (10)$$

Assuming that migration/error is distributed uniformly across the parallelogram shown in Figure 3a, the estimated population size on January 1st of each year is as follows:

$$P(x+n, t+n) = C_1(x) - \sum_{i=0}^{n-1} \left[D_U(x+i, t+i) + D_L(x+i+1, t+i) \right] + \frac{n}{5} \Delta_x \ , \quad (11)$$

for $n = 0, \ldots, 5$. By design, when $n = 0$ or 5 these population estimates match census counts exactly:

$$P(x, t) = C_1(x) \quad (12)$$

$$\text{and} \quad P(x+5, t+5) = C_2(x+5) \ . \quad (13)$$

2. New cohorts

The above formula applies only to cohorts who are already alive at the time of the first census. For cohorts born between the two censuses, intercensal population estimates are obtained by subtracting the number of deaths occurring before the second census from the number of births for the cohort. For a

cohort born in year $t+j$ within the intercensal interval, $[t, t+5)$, let

K = length of the interval $[t+j+1, t+5)$

= age (at last birthday) of the cohort born in year $t+j$ at the time of the second census

= $4-j$;

and B_{t+j} = number of births in year $t+j$.

An initial estimate of population size for the cohort born in year $t+j$ at the time of the second census is

$$\hat{C}_2(K) = B_{t+j} - D_L(0, t+j) - \sum_{i=1}^{K} \left[D_U(i-1, t+j+i) + D_L(i, t+j+i) \right], \quad (14)$$

and the difference between this estimate and the actual population count is

$$\Delta'_{t+j} = C_2(K) - \hat{C}_2(K) \ . \quad (15)$$

Thus, the estimated size of the cohort on January 1st of each year from birth until the second census is:

$$P(k, t+j+1+k) = B_{t+j} - D_L(0, t+j) - \sum_{i=1}^{k} \left[D_U(i-1, t+j+i) + D_L(i, t+j+i) \right] + \frac{2k+1}{2K+1} \Delta'_{t+j}, \quad (16)$$

for $k = 0, \ldots, K$. As before, population estimates at time $t+5$ match the counts in the second census exactly: $P(K, t+5) = C_2(K)$.

For example, consider the cohort born in year $t+2$. Thus, $j=2$, and $K = 4-j = 2$. In other words, the cohort born in year $t+2$ will be aged 2 at the time of the second census, as illustrated in Figure 3b. Population estimates for this cohort on January 1st of each year (until $t+5$) are as follows:

$$P(0, t+3) = B_{t+2} - D_L(0, t+2) + \frac{1}{5} \Delta'_{t+2}, \quad (17)$$

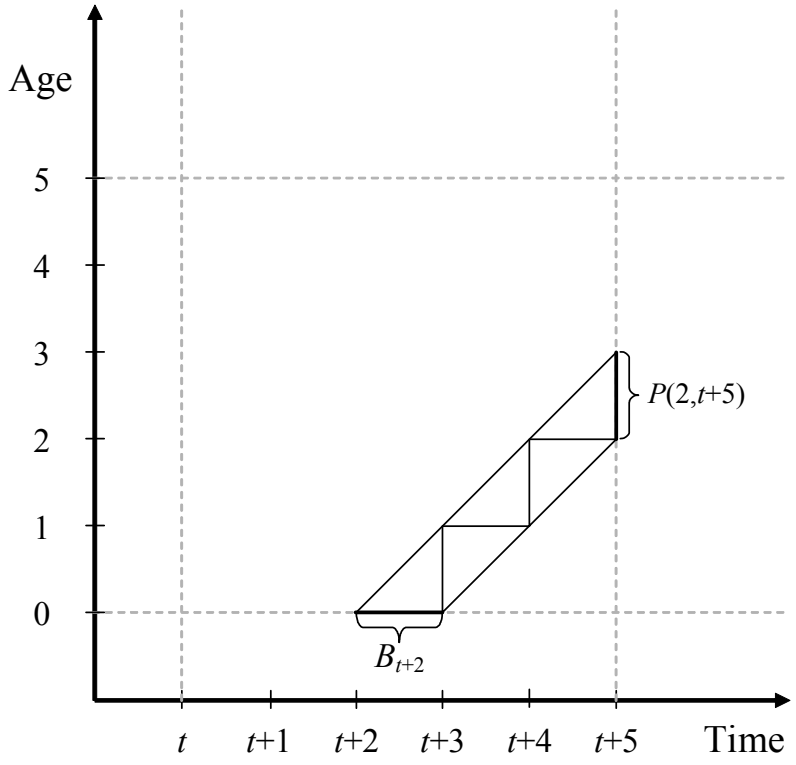
$$P(1, t+4) = B_{t+2} - D_L(0, t+2) - [D_U(0, t+3) + D_L(1, t+3)] + \frac{3}{5} \Delta'_{t+2}, \quad (18)$$

and

$$P(2, t+5) = B_{t+2} - D_L(0, t+2) - \sum_{i=1}^{2} \left[D_U(i-1, t+2+i) + D_L(i, t+2+i) \right] + \Delta'_{t+2} \ . \quad (19)$$

Figure 3. Intercensal survival method (example)

b) new cohorts



Generalizing the method

The arguments above make the explicit assumption that the two censuses bounding the intercensal period each occur on January 1st and are exactly five years apart. However, reality is typically more complicated. In this section, we generalize the method to allow for censuses that occur on any date of the year and for intercensal intervals of any length.

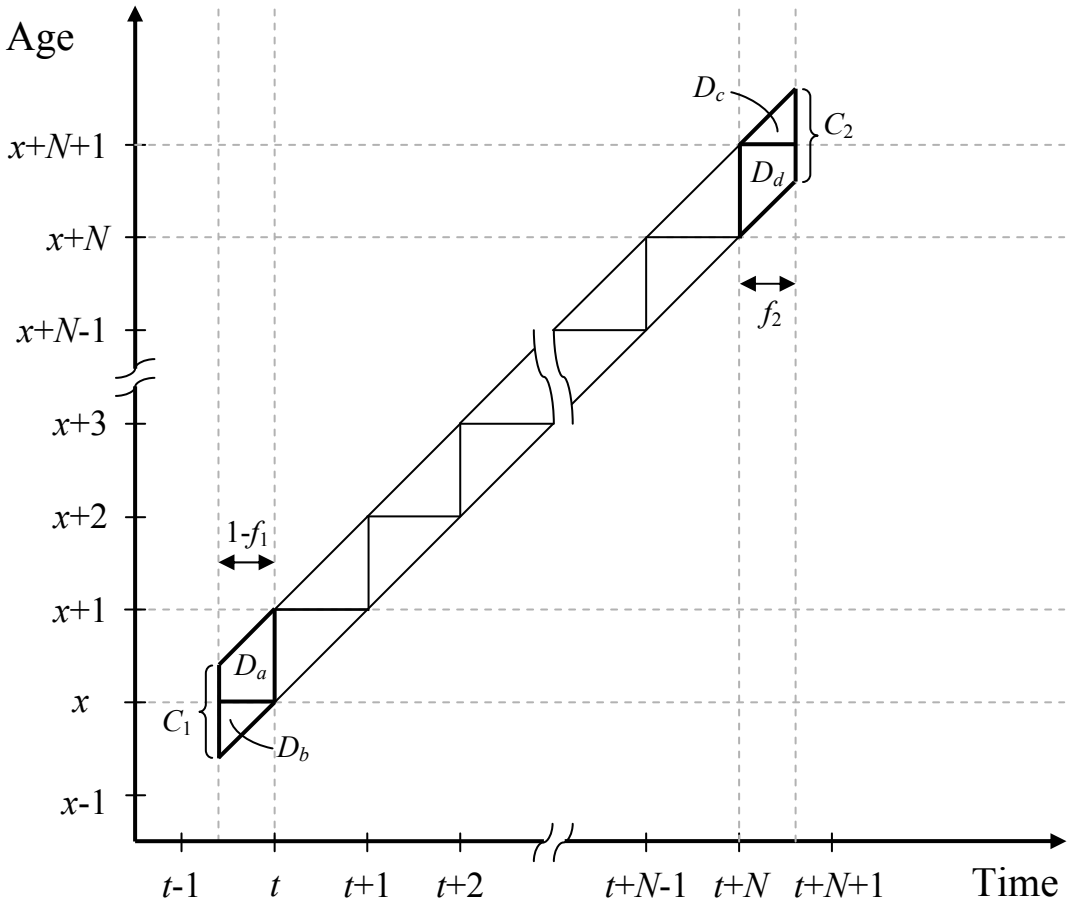
1. Pre-existing cohorts

Figure 4a depicts an intercensal period bounded by two censuses that occur on arbitrary dates. Let t and $t+N$ be the times of the first and the last January 1st within the intercensal interval. Thus, N equals the number of complete calendar years between the two censuses. Let f_1 be the fraction of calendar year $t-1$ before the first census, and let f_2 be the fraction of calendar year $t+N$ before the

second census. Thus, the two censuses occur at times $t_1 = t - 1 + f_1$ and $t_2 = t + N + f_2$, and the total length of the intercensal period is $N + 1 - f_1 + f_2$.

Figure 4. Intercensal survival method (in general)

a) pre-existing cohorts



The highlighted cohort in Figure 4a is of age x on January 1st of year t. This cohort was aged $x - 1$ or x at the time of the first census, and will be aged $x + N$ or $x + N + 1$ at the time of the second census. If individuals are uniformly distributed across their respective age intervals at each census enumeration, the sizes of this cohort at the beginning and end of the intercensal interval are

$$C_1 = (1 - f_1) \cdot C_1(x - 1) + f_1 \cdot C_1(x) \qquad (20)$$

and $$C_2 = (1 - f_2) \cdot C_2(x + N) + f_2 \cdot C_2(x + N + 1) \ , \qquad (21)$$

respectively. Although the assumption of a uniform distribution across age intervals is obviously incorrect, errors of exaggeration will tend to be balanced by those of understatement, yielding sufficiently accurate estimates in most cases.

Similarly, assuming a uniform distribution of deaths within Lexis triangles, deaths to this cohort in year $t-1$ after the first census enumeration will be composed of two components:

$$D_a = (1 - f_1^2) \cdot D_L(x, t-1) \quad (22)$$

and

$$D_b = (1 - f_1)^2 \cdot D_U(x-1, t-1) \ . \quad (23)$$

Likewise, under the same assumption, deaths to this cohort in year $t+N$ before the second census enumeration will be

$$D_c = f_2^2 \cdot D_L(x+N+1, t+N) \quad (24)$$

and

$$D_d = (2f_2 - f_2^2) \cdot D_U(x+N, t+N) \ . \quad (25)$$

Using these numbers along with death counts during complete calendar years of the intercensal interval, we estimate the size of the highlighted cohort at the time of the second census as follows:

$$\hat{C}_2 = C_1 - (D_a + D_b) - \sum_{i=0}^{N-1} \left[D_U(x+i, t+i) + D_L(x+i+1, t+i) \right] - (D_c + D_d) \ . \quad (26)$$

The difference between the actual census count and this estimate, $\Delta_x = C_2 - \hat{C}_2$, represents the total intercensal migration/error for this cohort. Finally, the size of the cohort on each January 1st of the intercensal interval is estimated as follows:

$$P(x+n, t+n) = C_1 - (D_a + D_b) - \sum_{i=0}^{n-1} \left[D_U(x+i, t+i) + D_L(x+i+1, t+i) \right] + \frac{1 - f_1 + n}{N + 1 - f_1 + f_2} \Delta_x \quad (27)$$

for $n = 0, \ldots, N$.

2. Infant cohort

The above formulas are applicable for cohorts that are aged 1 or more on the first January 1st of the intercensal interval. For the cohort aged 0 on this date (Figure 4b), and for new cohorts born during

the intercensal interval (Figure 4c), different formulas are needed. For the infant cohort, the following modifications to the above formulas are necessary:

$$C_1 = (1 - f_1) \cdot B_{t-1} + f_1 \cdot C_1(0) \ , \qquad (28)$$

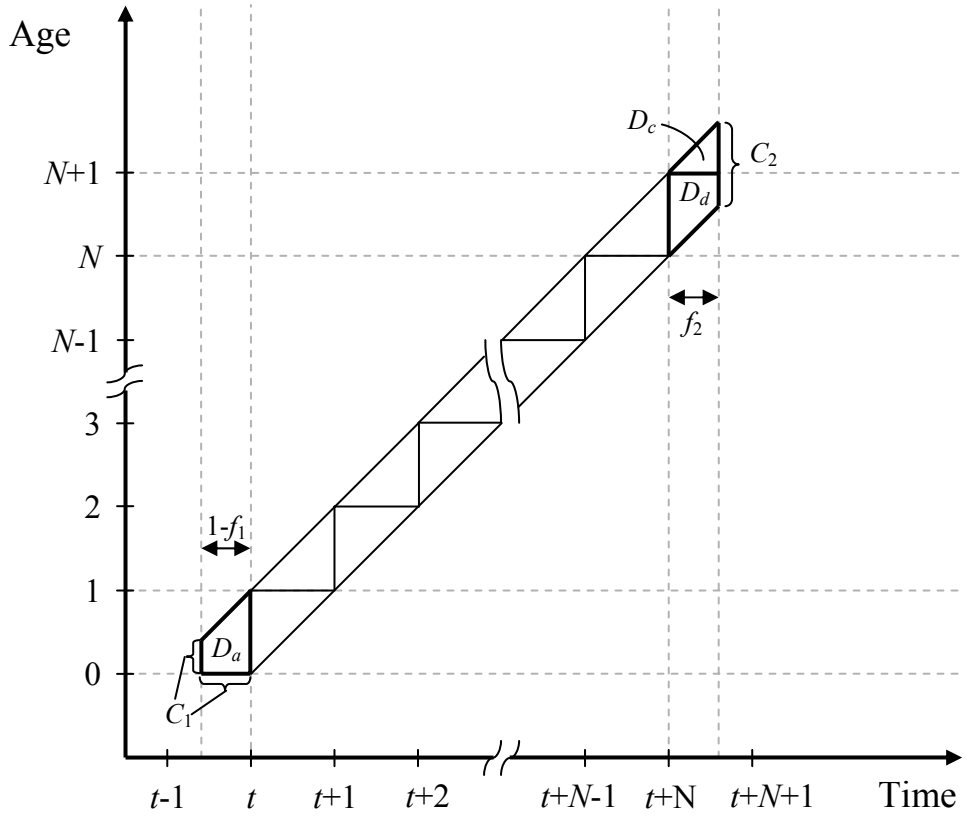
$$\hat{C}_2 = C_1 - D_a - \sum_{i=0}^{N-1} \left[D_U(i, t+i) + D_L(i+1, t+i) \right] - (D_c + D_d) \ , \qquad (29)$$

and $$P(n, t+n) = C_1 - D_a - \sum_{i=0}^{n-1} \left[D_U(i, t+i) + D_L(i+1, t+i) \right] + \frac{\frac{1}{2}\left(1 - f_1^2\right) + n}{N + \frac{1}{2}\left(1 - f_1^2\right) + f_2} \Delta_0 \ , \qquad (30)$$

for $n = 0, \ldots, N$, where $\Delta_0 = C_2 - \hat{C}_2$. Note the following four differences between these formulas and those given earlier: (1) x disappears from the latter two equations since $x = 0$; (2) in the first formula, $C_1(x-1)$ is replaced by B_{t-1}, the number of births during the calendar year of the first census; (3) in the latter two formulas, D_b is absent as it is undefined; and (4) in the last term of the third equation, $1 - f_1$ is replaced by $\frac{1}{2}\left(1 - f_1^2\right)$ in both numerator and denominator. The formulas for D_a, D_c, D_d, and C_2 are unaltered.

Figure 4. Intercensal survival method (in general)

b) infant cohort



3. New cohorts

Lastly, we consider the case of a cohort born during complete calendar years of the intercensal interval. A cohort born in year $t+j$ will be aged $K = N - j - 1$ on the last January 1st before the second census. Defining f_2, D_c, and D_d as before, the following equations are used to estimate the size of new cohorts on January 1st of each year (from birth until just before the second census):

$$\hat{C}_2 = B_{t+j} - D_L(0, t+j) - \sum_{i=1}^{K} \left[D_U(i-1, t+j+i) + D_L(i, t+j+i) \right] - (D_c + D_d) \ , \qquad (31)$$

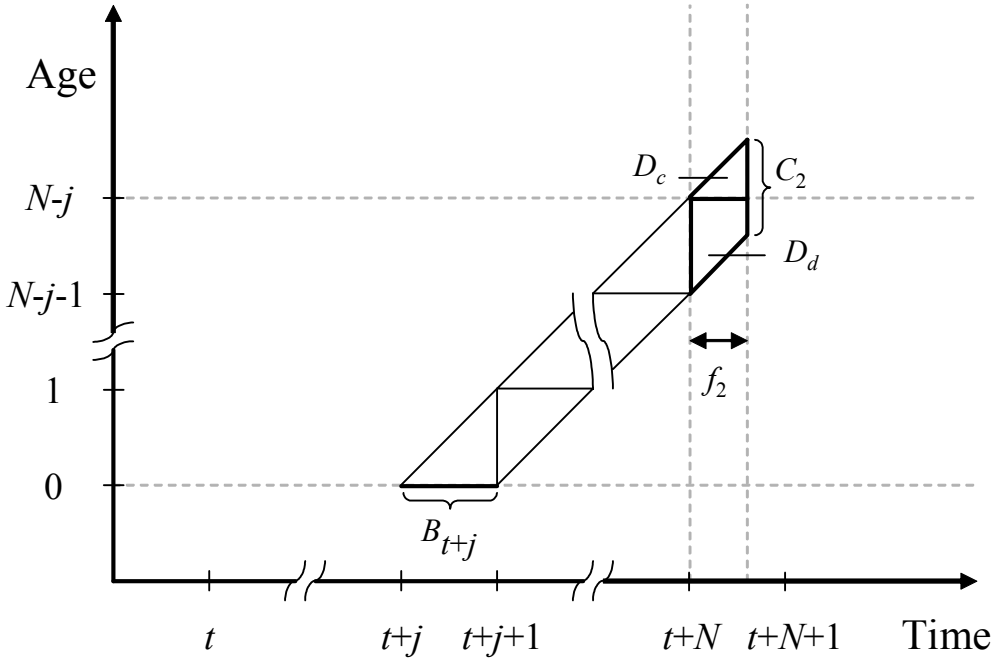
and

$$P(k, t+j+k+1) = B_{t+j} - D_L(0, t+j) - \sum_{i=1}^{k} \left[D_U(i-1, t+j+i) + D_L(i, t+j+i) \right] + \frac{2k+1}{2K+1+2f_2} \Delta'_{t+j} \ , \qquad (32)$$

for $k = 0, \ldots, K$, where $\Delta'_{t+j} = C_2 - \hat{C}_2$.

Figure 4. Intercensal survival method (in general)

c) new cohorts



Pre- and postcensal survival method

For a short period before the first census or after the last census, population estimates can be derived simply by adding or subtracting deaths from population counts in a census (or, for new cohorts, from birth counts). The formulas are similar to those presented earlier, although they lack a correction for migration/error. Therefore, population estimates for recent years that are derived in this manner must be considered provisional. They will be replaced by final estimates once another census is available to close the intercensal interval. The purpose of such estimates is to allow mortality estimation during recent years or for a short period before an early census, when appropriate death counts are available during an open census interval.

Examples of pre- and postcensal survival estimation are shown in Figure 5. The size of the cohort born in year $t-x-1$ on January 1st of years $t-1$ and $t-2$ is estimated as follows:

$$P(x-1,t-1) = C_1 + D'_a + D'_b \qquad (33)$$

and $$P(x-2,t-2) = C_1 + D'_a + D'_b + D_L(x-1,t-2) + D_U(x-2,t-2)\ . \qquad (34)$$

To estimate the size of the same cohort on January 1st of years $t+N+1$ and $t+N+2$, we have:

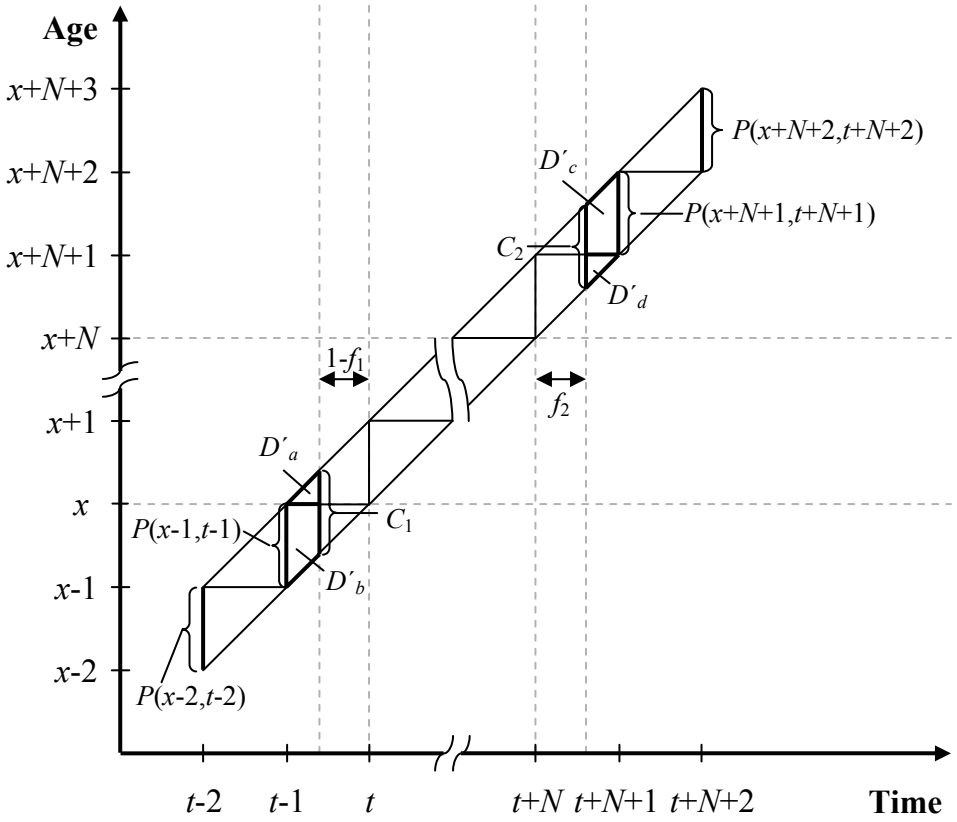
$$P(x+N+1,t+N+1) = C_2 - D'_c - D'_d \quad (35)$$

and

$$P(x+N+2,t+N+2) = C_2 - D'_c - D'_d - D_U(x+N+1,t+N+1) - D_L(x+N+2,t+N+1). \quad (36)$$

In this notation, D'_a, D'_b, D'_c, and D'_d, are the complements of D_a, D_b, D_c, and D_d, respectively. That is, the sum of each pair of death counts equals the number of deaths in a Lexis triangle. For example, comparing Figures 4a and 5, we see that $D'_a + D_a = D_L(x,t-1)$.

Figure 5. Pre- and postcensal survival method



Intercensal survival with census data in n-year age groups

The above discussion assumes that census data are available in single-year age groups. However, for many historical censuses the available counts refer to *n*-year age groups, where *n* is often 5. In these cases, we must first split the data into one-year age groups before computing population estimates using

the method of intercensal survival. We employ a simple method for this purpose. We assume that a more recent census is available, which contains population counts by single years of age. Using the age distribution at the time of the later census, plus death counts in the intercensal interval, we estimate the age distribution of the earlier census by the method of reverse survival. However, these estimates may not sum to the total (or sub-totals) given in the earlier census. Therefore, we use only the estimated distribution of the population by age at the time of the earlier census, which is applied to the observed counts within *n*-year age intervals as a means of creating finer age categories. Thus, all counts contained in the earlier census are preserved in the process of making these calculations.

Extinct cohorts methods

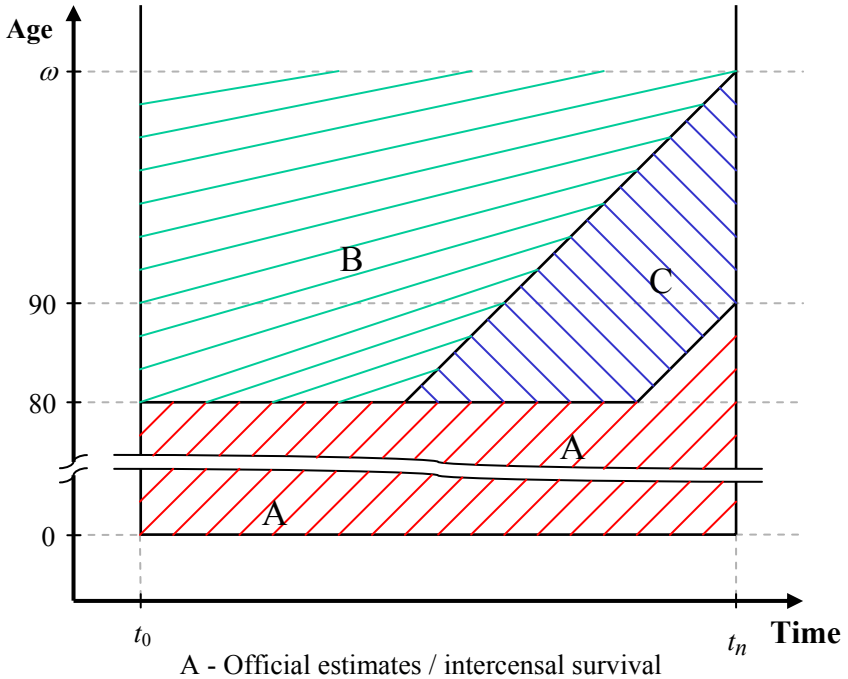
The method of extinct generations can be used to obtain population estimates for cohorts with no surviving members at the end of the observation period. With this method, the population size for a cohort at age *x* is estimated by summing all future deaths for the cohort, which can be written as follows:

$$P(x,t) = \sum_{i=0}^{\infty} \left[D_U(x+i, t+i) + D_L(x+i+1, t+i) \right] . \qquad (37)$$

This method assumes that there is no international migration after age *x* for the cohort in question, which is a reasonable assumption only for advanced ages. We use the method of extinct generations to estimate population sizes for ages 80 and above only, as illustrated in Figure 6.

Prior to applying the method of extinct cohorts, it is necessary to determine which cohorts are extinct. For this purpose, we adopt a method proposed by Väinö Kannisto and used already in the Kannisto-Thatcher oldest-old mortality database (Andreev, 2001). We say that a cohort is extinct if it has attained age ω by end of the observation period (assumed to occur on January 1st of year t_n). Thus, we need to find ω or, equivalently, ω-1, the age of the oldest non-extinct cohort.

Figure 6. Methods used for population estimates



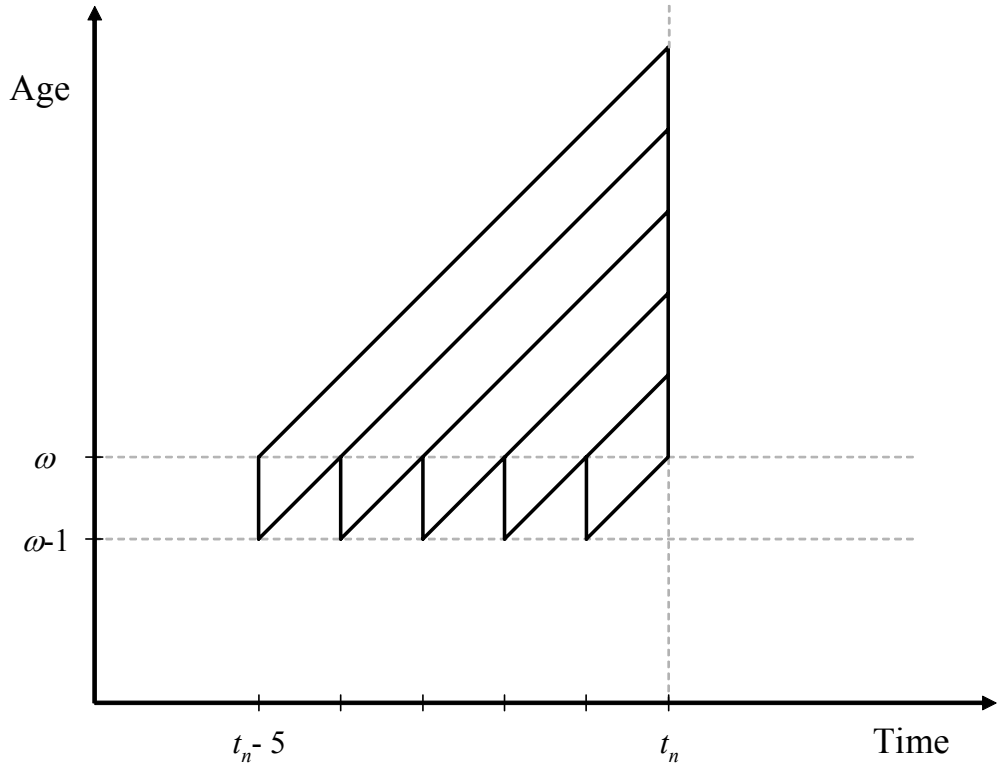
A - Official estimates / intercensal survival
B - Extinct cohorts
C - Survivor ratio, SR90+

Consider a cohort aged x at the end of the observation period, where x is some very high age (like 120). We examine the most recent ℓ cohorts from a similar point in their life histories. Specifically, we consider the observed deaths for these cohorts from January 1st of the year when they were aged x until the end of the observation period (see illustration in Figure 7, where $\ell = 5$ and $x = \omega - 1$). For these cohorts over the specified intervals of age and time, we compute the average number of deaths:

$$\widetilde{D}(x, t_n, \ell) = \frac{1}{\ell} \sum_{j=1}^{\ell} \sum_{i=0}^{j-1} \left[D_U(x+i, t_n - j + i) + D_L(x+i+1, t_n - j + i) \right], \qquad (38)$$

with $\ell = 5$. For very high ages, $\widetilde{D}(x, t_n, \ell)$ will be close to zero. We define ω to be the lowest age x such that $\widetilde{D}(x, t_n, \ell) \le 0.5$. Equivalently, ω-1 is the highest age x for which $\widetilde{D}(x, t_n, \ell) > 0.5$.

Figure 7. Illustration of extinction rule (with $\ell = 5$ and $x = \omega - 1$)



Survivor ratio

The survivor ratio method is used to estimate population sizes above age 80 for almost-extinct cohorts (see Figure 6). The method is applied to cohorts that are at least age 90 at the end of the observation period but not yet extinct (according to the rule given above).[16] Various versions of this method have been proposed and studied previously (see discussion in Andreev, 1999). We use the version that proved most reliable in an earlier comparative study (Thatcher *et al.*, 2002).

Define a "survivor ratio" to be the ratio of survivors alive at age x on January 1st of year t to

[16] We make an exception for the small number of countries that have reliable January 1st population estimates by single year of age to the maximum age ω for the last year of observation (i.e., Sweden, Denmark, Norway, Finland, and Iceland). For these countries, we use the official population estimates for ages 90 and older on January 1st of year t_n and derive population estimates in earlier years (for each cohort) by adding observed death counts back to age 80 (like for the extinct cohort method).

those in the same cohort who were alive *k* years earlier:

$$R = \frac{P(x,t)}{P(x-k,t-k)} \ . \qquad (39)$$

Assuming that there is no migration in the cohort over the interval, this ratio can also be written:

$$R = \frac{P(x,t)}{P(x,t)+\dot{D}} \ , \qquad (40)$$

where $\dot{D} = \sum_{i=1}^{k} \left[D_U(x-i,t-i) + D_L(x-i+1,t-i) \right]$. Solving this equation for $P(x,t)$, we obtain:

$$P(x,t) = \frac{R}{1-R}\dot{D} \ . \qquad (41)$$

The survivor ratio for the oldest non-extinct cohort (aged ω-1 at time t_n) is illustrated in Figure 8. This survivor ratio is unknown, since we do not know the size of the cohort, $P(\omega-1,t_n)$, at the end of the observation period. However, comparable survivor ratios (i.e., with age ω-1 in the numerator) for all previous cohorts are available, since population size can be estimated using the method of extinct cohorts.

Suppose that a survivor ratio has approximately the same value for the cohort in question and for the previous *m* cohorts. That is, suppose that

$$R(x,t,k) = \frac{P(x,t)}{P(x-k,t-k)} \approx \frac{P(x,t-1)}{P(x-k,t-k-1)} \approx \cdots \approx \frac{P(x,t-m)}{P(x-k,t-k-m)} \ . \qquad (42)$$

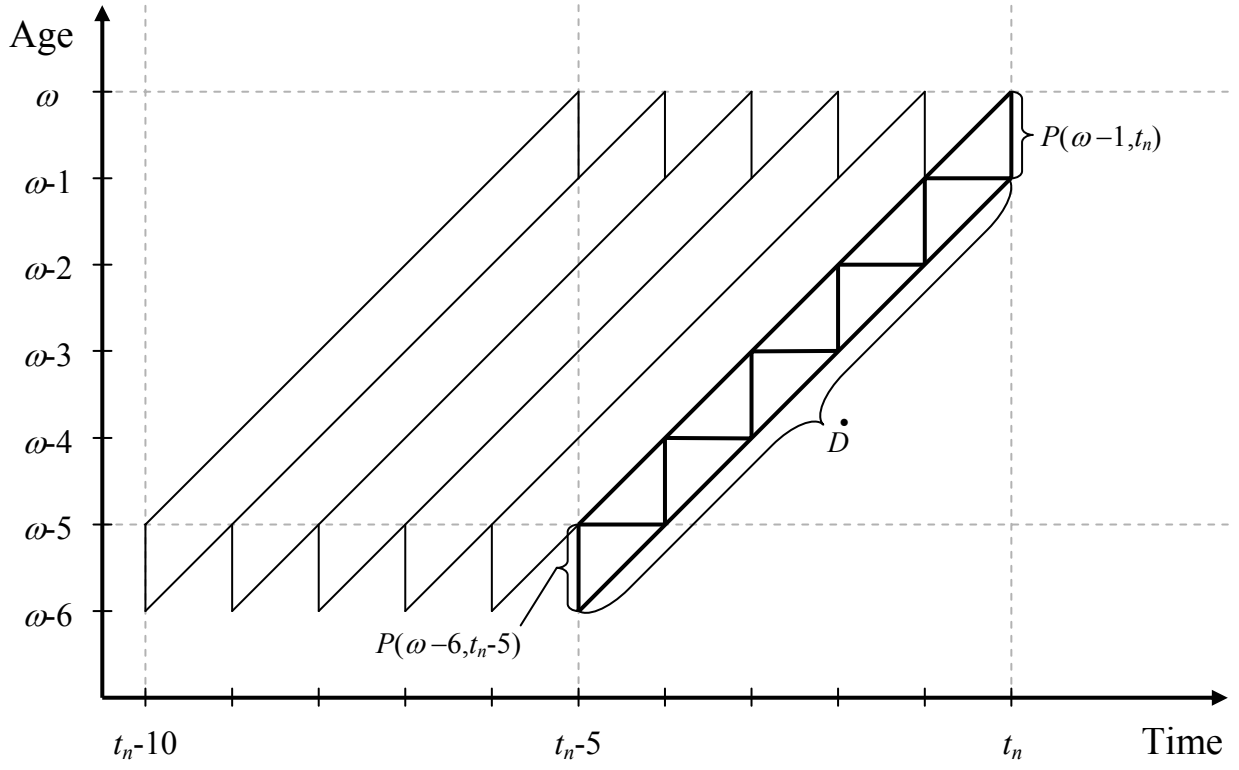
Then, we can estimate *R* by computing the pooled survivor ratio for the *m* previous cohorts:

$$R^*(x,t,k) = \frac{\sum_{i=1}^{m} P(x,t-i)}{\sum_{i=1}^{m} P(x-k,t-k-i)} \ . \qquad (43)$$

If both R^* and $\dot{D}$ are available for a given cohort, we can estimate $P(x,t)$ as follows:

$$\widetilde{P}(x,t) = \frac{R^*}{1-R^*}\dot{D} \ . \qquad (44)$$

Figure 8. Survivor ratio method (at age $x = \omega - 1$, with $k = m = 5$)



In the simplest version of the survivor ratio method, this procedure is used to obtain $P(\omega-1,t_n)$, and then the size of this cohort in previous years is estimated by adding observed death counts back to age 80 (in a fashion similar to the extinct cohort method). It is then possible to apply the same method recursively to obtain $P(\omega-2,t_n)$, $P(\omega-3,t_n)$, etc., down to some lower age limit (e.g., 90 years). This method works well if its fundamental assumption is not violated, that is, if the survivor ratios for successive cohorts are nearly equal. A common occurrence, however, is that these survivor ratios increase over time as a result of mortality decline. Therefore, R^* tends to underestimate R, and $\widetilde{P}$ tends to underestimate P.

These considerations motivate a modified version of the survivor ratio estimate:

$$\begin{aligned}\hat{P}(x,t) &= c\,\widetilde{P}(x,t) \\ &= c\frac{R^*}{1-R^*}\dot{D} \quad ,\end{aligned} \qquad (45)$$

where *c* is a constant that must be estimated. When mortality is declining/increasing/constant, *c* should be greater than/less than/equal to one. The problem, obviously, is how to choose the proper value of *c*.

Following Thatcher *et al.* (2002), we choose a value of *c* such that

$$\sum_{x=90}^{\omega-1} \hat{P}(x,t_n) = P(90+,t_n) \ , \qquad (46)$$

where $P(90+,t_n)$ is an official estimate of the population size in the open interval aged 90 and above at the end of the observation period. This version of the survival ratio method is known as SR(90+) and is used for the HMD (with $k = m = 5$) in all cases where $P(90+)$ is available and is believed to be reliable.[17] Otherwise, we use the simpler version of the survival ratio method (i.e., with $c = 1$).

Death rates

Death rates consist of death counts divided by the exposure-to-risk. In the case of a one-year age group and a single calendar year (i.e., a 1x1 period death rate), we have the following formula:

$$M_{xt}^{p} = \frac{D_{xt}^{p}}{E_{xt}^{p}} \ , \qquad (47)$$

where

$$D_{xt}^{p} = D_L(x,t) + D_U(x,t) \qquad (48)$$

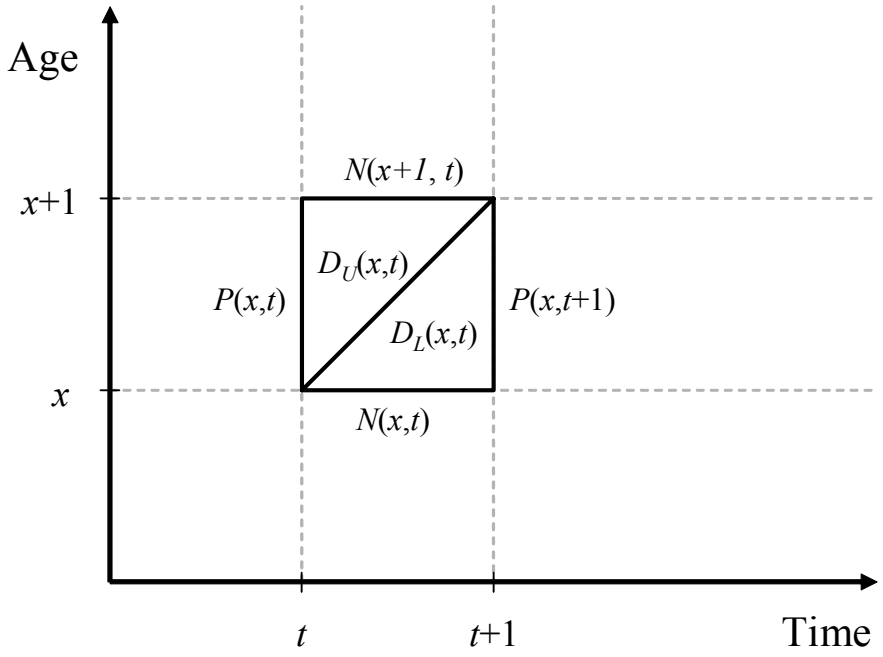
and E_{xt}^{p} is the exposure-to-risk in the age interval $[x, x+1)$ during calendar year *t*. The exposure-to-risk is always measured in terms of person-years and, for periods, is computed by the following formula (see Appendix E for a derivation):

$$E_{xt}^{p} = \frac{1}{2}\left[P(x,t) + P(x,t+1)\right] + \frac{1}{6}\left[D_L(x,t) - D_U(x,t)\right] . \qquad (49)$$

The data used for computing these quantities are illustrated in Figure 9.

[17] For some populations, official population estimates are available only for age 85+. In such cases, we use SR(85+) and note this modification in the general comments (see country-specific documentation for details).

Figure 9. Data for period death rates and probabilities



Cohort formulas are only slightly different. A 1x1 cohort death rate is

$$M_{xt}^{c} = \frac{D_{xt}^{c}}{E_{xt}^{c}} \ , \tag{50}$$

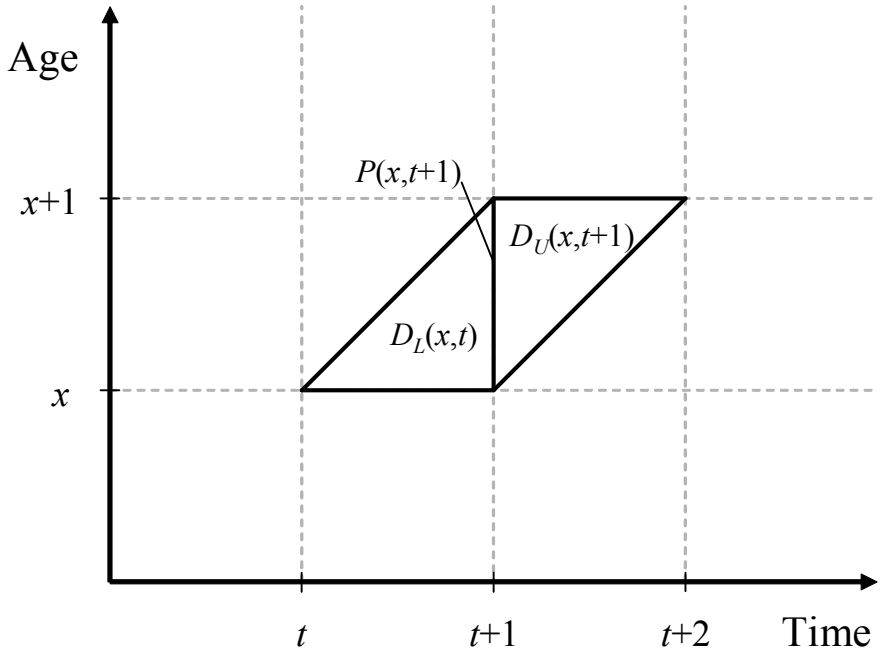
where death counts and exposure estimates (except at age 0) are defined as follows:

$$D_{xt}^{c} = D_{L}(x,t) + D_{U}(x,t+1) \tag{51}$$

$$E_{xt}^{c} = P(x,t+1) + \frac{1}{3}\left[D_{L}(x,t) - D_{U}(x,t+1)\right] \ . \tag{52}$$

The data used for computing these quantities are illustrated in Figure 10. The exposure estimate at age 0 is an exception. Because the cohort life table death rate m_{0t}^{c} is derived differently at age 0 than at other ages (see p.40), we define $E_{0t}^{c} = \frac{D_{0t}^{c}}{m_{0t}^{c}}$ in order to ensure that $M_{0t}^{c} = m_{0t}^{c}$.

Figure 10. Data for cohort death rates and probabilities



For broader intervals of age and/or time (whether time is defined by periods or cohorts), death rates are always found by pooling deaths and exposures first and then dividing the former by the latter. Throughout the rest of this discussion, we will refer either to one-year or five-year death rates (i.e., M_x or ${}_5M_x$). For simplicity of notation, we will not specify a particular time interval, because the formulas for computing probabilities of death and/or life tables are the same for any interval of time. Also, we will not use a 'p' or 'c' superscript to distinguish between period and cohort death rates, since the difference should always be apparent from the context.

Life tables

Life table calculations do not depend on the organization of the data over time. For any time interval, the same methods are used for computing life tables from a set of age-specific death rates. However, the methods used here are slightly different for period and cohort life tables. Period tables are computed by converting death rates to probabilities of death. Before this conversion, death rates at older ages are smoothed by fitting a logistic function. For cohort life tables, we compute probabilities of death directly from the data and perform no smoothing at older ages. As discussed in Appendix E, cohort

probabilities of death computed in this manner are fully consistent with the cohort death rates described in the previous section.

For both periods and cohorts, we begin by computing "complete" life tables (i.e., single-year age groups) using our final estimates of death counts by Lexis triangle and population size (on January 1st) by single years of age. Then, "abridged" tables (i.e., five-year age groups) are extracted from the complete tables. Deriving abridged tables from complete ones (rather than computing them directly from data in five-year age intervals) ensures that both sets of tables contain identical values of life expectancy and other quantities.

Period life tables

Whereas a cohort life table depicts the life history of a specific group of individuals, a period life table is supposed to represent the mortality conditions at a specific moment in time. However, observed period death rates are only one result of a random process for which other outcomes are possible as well. At older ages where this inherent randomness is most noticeable, it is well justified to smooth the observed values in order to obtain an improved representation of the underlying mortality conditions. Thus, for period life tables by single years of age, we begin by smoothing observed death rates at older ages by fitting a logistic function to observed death rates for ages 80 and above, separately for males and females.[18]

Suppose that we have deaths D_x and exposure E_x for ages $x = 80$, 81, … , 110+ (for convenience, we define x=110 for the open category above age 110). We smooth observed death rates $M_x = \frac{D_x}{E_x}$ by fitting the Kannisto model of old-age mortality (Thatcher et al., 1998), with an asymptote

[18] It is a common actuarial practice to fit a curve to death rates at older ages in the process of computing a life table. We use the logistic function because a recent study concluded that such a curve fits the mortality pattern at old ages at least as well as, and usually better than, any other mortality model (Thatcher *et al.*, 1998). Fixing the value of the asymptote at one simplifies these calculations and avoids certain anomalies that may occur as a result of random fluctuations. In any event, estimates of this asymptote have been around one in most previous studies.

equal to one, to estimate the underlying hazards function μ_x:[19]

$$\mu_x(a,b) = \frac{ae^{b(x-80)}}{1+ae^{b(x-80)}} \ , \qquad (53)$$

where we require $a \geq 0$ and $b \geq 0$.[20] Assuming that $D_x \sim Poisson\left(E_x \mu_{x+0.5}(a,b)\right)$, we derive parameter estimates $\hat{a}$ and $\hat{b}$ by maximizing the following log-likelihood function:

$$\log L(a,b) = \sum_{x=80}^{110} \left[D_x \log \mu_{x+0.5}(a,b) - E_x \mu_{x+0.5}(a,b)\right] + \text{constant} \ . \qquad (54)$$

Substituting $\hat{a}$ and $\hat{b}$ into equation (53) yields smoothed death rates $\hat{M}_x$, where $\hat{M}_x = \hat{\mu}_{x+0.5} = \mu_{x+0.5}(\hat{a},\hat{b})$. In this model specification $\hat{a}$ and $\hat{b}$ are constrained to be positive so that smoothed death rates cannot decline above age 80. For the rest of the calculations described here, fitted death rates replace observed death rates for all ages at or above *Y*, where *Y* is defined as the lowest age where there are fewer than 100 deaths, but is constrained to $80 \leq Y \leq 95$.[21] Thus, complete period life tables for males and females are constructed based on the following vector of death rates: M_0, M_1, … , M_{Y-1}, $\hat{M}_Y$, … , $\hat{M}_{109}$, ${}_{\infty}\hat{M}_{110}$.

After obtaining smoothed death rates for males and females, we calculate the smoothed rates for the total population as a weighted average of those for males and females:

$$\hat{M}_x^T = w_x^F \hat{M}_x^F + (1 - w_x^F)\hat{M}_x^M \ , \qquad (55)$$

[19] This smoothing procedure is used only if there are at least two $M_x > 0$ at ages 80 and older. If there are fewer than two non-zero observed death rates, then we assume the death rate is constant for ages $[x_{\varpi}, 110]$, where x_{ϖ} is the oldest age where $M_x > 0$.

[20] In order to satisfy the constraints on the parameters *a* and *b*, we fit the model in terms of *a** and *b**, where $a = e^{a^*}$ and $b = e^{b^*}$.

[21] In other words, we used the fitted death rates for all ages at or above the greater of 80 or the lowest age where there are fewer than 100 *male* deaths (because at older ages there are typically fewer male deaths than female deaths), and for all ages at or above age 95 regardless of the number of deaths. We begin using fitted death rates at the same age for both males and females.

where superscripts T, F, and M represent total, female, and male, respectively, and w_x^F represents the weight for females aged *x*, but these must still be determined.

For observed death rates, the analogous weights equal the observed the proportion of female exposure:

$$\pi_x^F = \frac{E_x^F}{E_x^F + E_x^M} = \frac{E_x^F}{E_x^T} \quad . \tag{56}$$

For smoothed rates as well, such weights could be calculated from observed exposures, but due to random fluctuations in such values at older ages, the resulting series of death rates for the total population would not be as smooth as those for males and females. Consequently, we smooth π_x^F itself by fitting the following model by the method of weighted least squares:

$$z = \text{logit}(\pi_x^F) = \ln\left(\frac{\pi_x^F}{1-\pi_x^F}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 \quad . \tag{57}$$

We drop observations where E_x^F, E_x^M, or both equal 0 (in such cases, π_x^F=0 or 1 and thus the logit is undefined), and for fitting equation 57, use weights equal to E_x^T.[22] The fitted values are obtained as follows:

$$\hat{z} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \quad \text{and} \quad w^F = \hat{\pi}_x^F = \frac{e^{\hat{z}}}{1+e^{\hat{z}}} \quad . \tag{58}$$

Finally, the smoothed total death rates are calculated as:

$$\hat{M}_x^T = \hat{\pi}_x^F \hat{M}_x^F + (1-\hat{\pi}_x^F)\hat{M}_x^M \quad . \tag{59}$$

We then assume that death rates in the life table equal death rates in the population (i.e., that $m_x = M_x$ for ages $x = 0$, 1, …, *Y-1*, that $m_x = \hat{M}_x$ for *x* = *Y*, *Y*+1, …, 109, and that ${}_\infty m_{110} = {}_\infty \hat{M}_{110}$ for the

[22] For fitting the model in equation 57, theoretically, the correct weights would be $\hat{\pi}_x^F(1-\hat{\pi}_x^F)\cdot E_x^T$, but would require an iterative procedure because the weights depend on the fitted values themselves. Since there is relatively little variability in $\hat{\pi}_x^F(1-\hat{\pi}_x^F)$ compared to E_x^T over the observed range, using E_x^T as the weights should provide reasonable accuracy and is much more convenient.

open age interval above 110). This assumption is correct only when the age structure of the actual population is identical to the age structure of a stationary (i.e., life table) population within each age interval (for more explanation, see Keyfitz, 1985, or Preston *et al.*, 2001). In most situations, however, deviations from this assumption are likely to be small and unimportant for one-year age intervals.

Next, we convert the life table death rates, m_x, into probabilities of death, q_x. Let a_x be the average number of years lived within the age interval $[x, x+1)$ for people dying at that age. We assume that $a_x = \frac{1}{2}$ for all single-year ages except age 0 (see below). We then compute q_x from m_x and a_x according to the formula,

$$q_x = \frac{m_x}{1 + (1 - a_x) \cdot m_x} \quad , \tag{60}$$

for $x = 0$, 1, 2, …, 109. For the open age interval, we set ${}_{\infty}a_{110} = \frac{1}{{}_{\infty}m_{110}}$ and ${}_{\infty}q_{110} = 1$.

For infants, we adopt the formulas for a_0 suggested by Preston *et al.* (2001: 48), which are adapted from the Coale-Demeny model life tables (Coale and Demeny, 1983). Thus, if $m_0 \geq 0.107$:

$$a_0 = \begin{cases} 0.350 & \text{for females} \\ 0.330 & \text{for males} \end{cases} \quad . \tag{61}$$

On the other hand, if $m_0 < 0.107$:

$$a_0 = \begin{cases} 0.053 + 2.800 \cdot m_0 & \text{for females} \\ 0.045 + 2.684 \cdot m_0 & \text{for males} \end{cases} \quad . \tag{62}$$

For a life table with sexes combined, we compute a_0 as follows:

$$a_0^T = \frac{a_0^F D_0^F + a_0^M D_0^M}{D_0^F + D_0^M} \quad , \tag{63}$$

where the superscripts *F*, *M*, and *T* denote values for the female, male, and total populations, respectively, and where D_0^i refers to all deaths at age zero (both lower and upper triangles) for population *i*.

To complete the life table calculation, let p_x be the probability of surviving from age x to $x+1$. Therefore,

$$p_x = 1 - q_x \ , \tag{64}$$

for all ages x. Let the radix of the life table be $l_0 = 100{,}000$. Then, the number of survivors (out of 100,000) at age x is

$$l_x = l_0 \cdot \prod_{i=0}^{x-1} p_i \ . \tag{65}$$

The distribution of deaths by age in the life-table population is

$$d_x = l_x \cdot q_x \ , \tag{66}$$

for $x = 0$, 1, …, 109. For the open age category, ${}_{\infty}d_{110} = l_{110}$.

The person-years lived by the life-table population in the age interval $[x, x+1)$ are

$$L_x = l_x - (1 - a_x) \cdot d_x \ , \tag{67}$$

for $x = 0$, 1, …, 109. For the open age category, ${}_{\infty}L_{110} = l_{110} \cdot a_{110}$. The person-years remaining for individuals of age x equal

$$T_x = \sum_{i=x}^{109} L_i + {}_{\infty}L_{110} \tag{68}$$

for $x = 0$, 1, …, 109. For the open age category, $T_{110} = {}_{\infty}L_{110}$. Remaining life expectancy at age x is

$$\mathring{e}_x = \frac{T_x}{l_x} \ , \tag{69}$$

for $x = 0$, 1, …, 110.

Cohort life tables

We now describe the method used to compute q_x, a_x, and m_x for cohorts. These are found using death counts in Lexis triangles (whether observed in this format or estimated from 1x1 or 5x1 data, as described above) and population estimates by single years of age and time. Once these values are available, a complete cohort life table is calculated using the same formulas as in the case of period tables.

The method for multi-year cohorts is similar, as described below.

The probability of dying within a year for the cohort who turns age *x* one year and age $x+1$ the following year is

$$q_x = 1 - \frac{P - D_U}{P + D_L} = \frac{D_L + D_U}{P + D_L} \ , \qquad (70)$$

where $P = P(x, t+1)$ is the size of the cohort in between the two calendar years (i.e., on January 1st of the latter year), $D_L = D_L(x,t)$ refers to lower triangle deaths for the year the cohort turns age x, and $D_U = D_U(x, t+1)$ refers to upper triangle deaths in the following year (see Figure 10). This formula is exact in the absence of migration, and it is a reasonable approximation in most situations, assuming that the direction and magnitude of migration are similar in both upper and lower triangles (see Pressat, 1980).

If we assume once again that deaths are distributed uniformly within Lexis triangles, the number of years lived in the age interval $[x, x+1)$ by someone dying at that age is

$$a_x = \frac{\frac{1}{3} D_L + \frac{2}{3} D_U}{D_L + D_U} \ , \qquad (71)$$

as shown in Appendix E. A formula for the cohort death rate was given earlier, and in this case $m_x = M_x$ automatically (since the actual population is the same as the life table population). From these three formulas, it is easy to verify that q_x, m_x, and a_x are related as shown in equation (60). Thus, an advantage of the method used here is that q_x, m_x, and a_x obey this classic formula even though the three quantities are derived independently from death counts and population estimates.

Although this method works well in general, it should not be used at age 0, since the assumption of a uniform distribution of deaths is not true even approximately at this age. However, this assumption underlies the above method for computing a_x and m_x, but not q_x. Therefore, a special procedure is used for computing a_0 and m_0 from q_0, which is found by the same method used at other ages. Given q_0, a_0 and m_0 are derived by an iterative procedure that uses both the relationship implied by equation (60) and the Coale-Demeny formulas used for computing period life tables (see equations 61 and 62). Iteration is

required because the value of a_0 is specified in terms of m_0, not q_0. Thus, we assign starting values for the iterative procedure by setting $m_0^{(0)} = q_0$ and then computing $a_0^{(0)}$ from $m_0^{(0)}$ according to the Coale-Demeny equations. On subsequent iterations, $m_0^{(n)}$ is derived from q_0 and $a_0^{(n-1)}$ (which requires solving equation 60 for m_x), and then $a_0^{(n)}$ is computed from $m_0^{(n)}$ based on the Coale-Demeny formulas. This procedure converges quickly, usually after 3 or 4 iterations. Given the resulting estimate of m_0, we assume that $m_0 = M_0$ and define E_0 so that equation 50 holds true.

If some members of a cohort are still alive at age 110, the above formulas are used for ages $x = 0, 1, \ldots, 109$ only. In this situation, for the open interval above age 110, we set ${}_\infty q_{110} = 1$, ${}_\infty m_{110} = {}_\infty M_{110} = \frac{{}_\infty D_{110}}{{}_\infty E_{110}}$, and ${}_\infty a_{110} = \frac{1}{{}_\infty m_{110}}$. On the other hand, if the cohort dies out before age 110, the earlier definitions of q_x, m_x, and a_x are used up to and including the age of extinction (note that $q_x = 1$ in the final age group) and all values are marked as "missing" at higher ages.

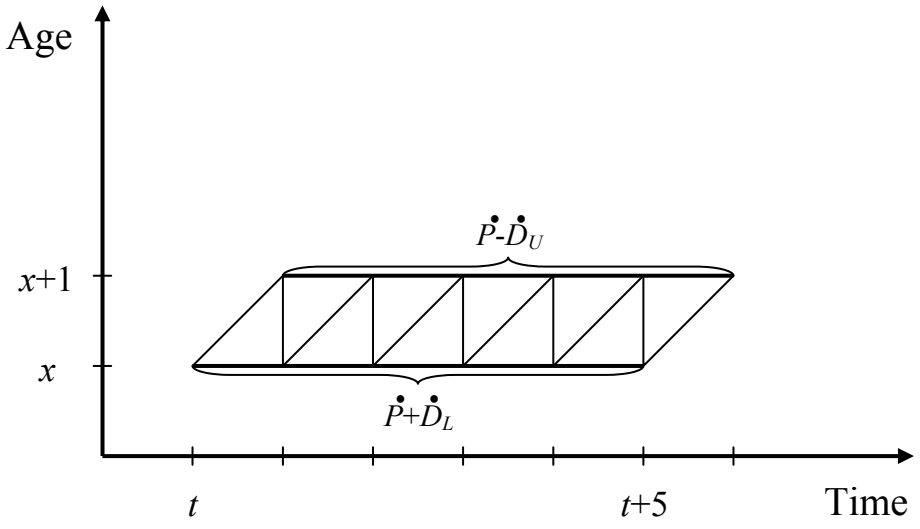
Multi-year cohorts

In the earlier section on death rates, we noted that death rates for multi-year intervals (either periods or cohorts) are found by pooling deaths and exposures first and then dividing the former by the latter. We now describe methods for computing cohort probabilities of death for time periods longer than one year using a similar principle (see Pressat, 1980).

Consider the example of an *n*-year birth cohort in the age interval from *x* to $x+1$. Let $\dot{P}$ denote the sum of the January 1st population estimates for the *n* individual birth cohorts when they are aged *x*. Likewise, let $\dot{D}_L$ and $\dot{D}_U$ denote the sums of lower and upper triangle deaths within the same age interval for the same group of cohorts (see Figure 11 for an illustration in the case where $n = 5$). Therefore, the probability of death for this *n*-year cohort is

$$q_x = 1 - \frac{\dot{P} - \dot{D}_U}{\dot{P} + \dot{D}_L} = \frac{\dot{D}_L + \dot{D}_U}{\dot{P} + \dot{D}_L} \quad . \qquad (72)$$

Figure 11. Illustration of five-year cohort (assuming no migration)



Using this notation, the death rate for this *n*-year cohort is

$$m_x = M_x = \frac{\dot{D}_L + \dot{D}_U}{\dot{P} + \frac{1}{3}\left(\dot{D}_L - \dot{D}_U\right)}, \qquad (73)$$

and the average age at death in the interval is

$$a_x = \frac{\frac{1}{3}\dot{D}_L + \frac{2}{3}\dot{D}_U}{\dot{D}_L + \dot{D}_U}. \qquad (74)$$

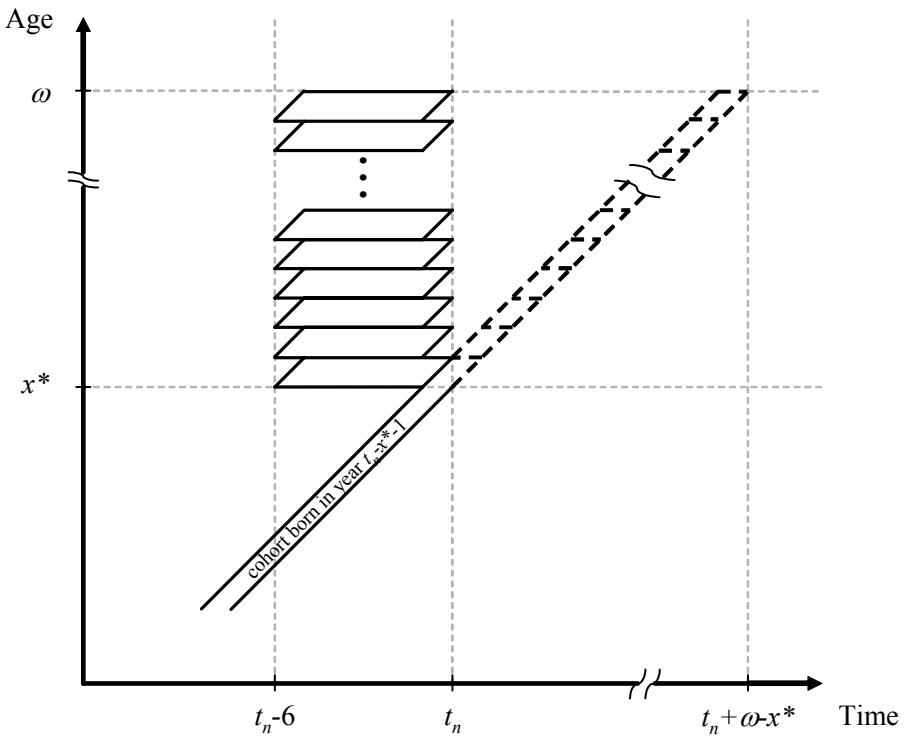
In this case it is easy to confirm that the relationship between these three quantities obeys the classical formula exactly. As with single-year cohorts, however, we compute a_0 and m_0 by the special iterative procedure described in the last section.

Almost-extinct cohorts

The above description assumes that all members of a cohort have died before we compute its life table. However, it is often desirable to compute life tables for cohorts that are almost extinct. Suppose that t_n denotes the last moment of the observation period and that the age of a cohort is x^* at time t_n. In order to compute a life table for this cohort, it is necessary to make some assumption about mortality at ages $x \geq x^*$. A simple solution is to assume that the cohort's deaths and exposures at these ages will

equal the average of those quantities for the most recent five-year cohort for whom such values are observed, as depicted schematically in Figure 12. Thus, the values of m_x, a_x, and q_x for an almost extinct cohort at ages $x \geq x^*$ are identical to those of a five-year cohort of comparable age observed just before time t_n.

Figure 12. Life table calculations for almost-extinct cohorts

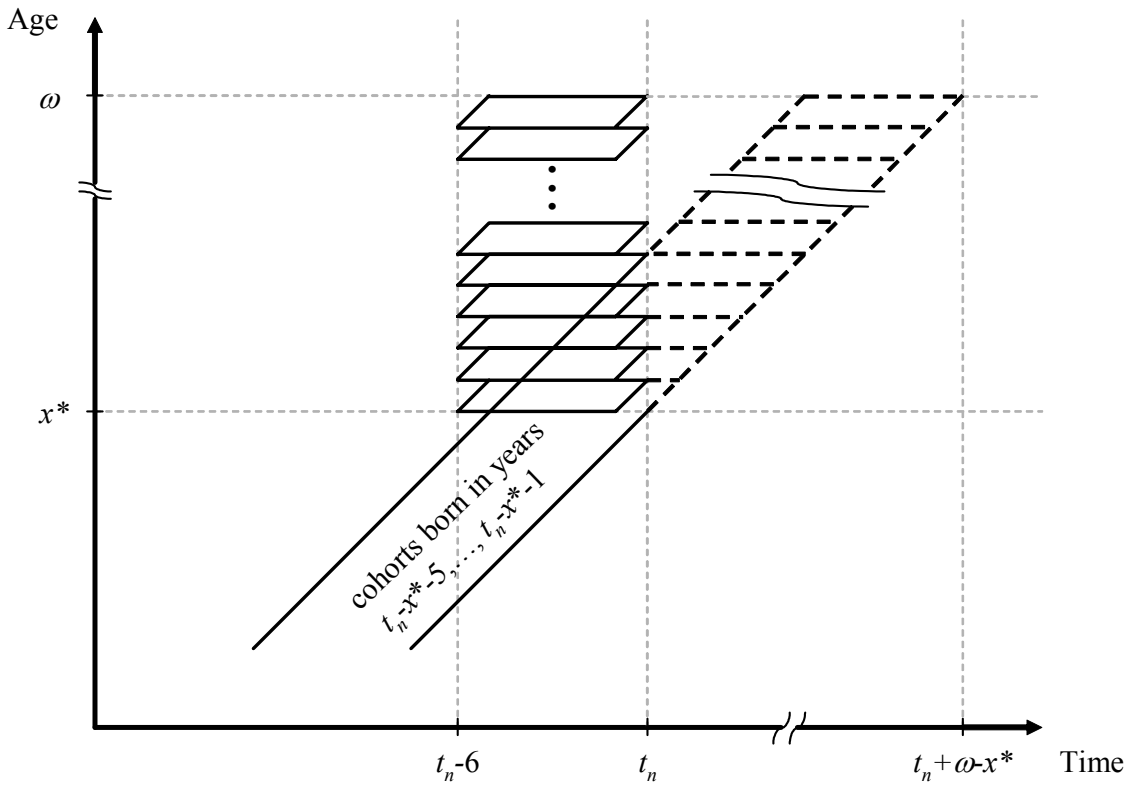


It is important to define some minimal value of x^* that is acceptable when making such calculations. For life tables that begin at age 0, we require that the total fictitious exposure (for ages $x \geq x^*$) be no more than one percent of the total lifetime exposure (in person-years lived) for the cohort. For life tables that begin at some age above 0, the fictitious exposure should be no more than one percent of the total exposure above the starting age. Figure 13 depicts life table calculations for five birth cohorts aged x^* to $x^* + 4$ at time t_n.

A small problem is that this method may produce $q_x = 1$ at some high age, even though there are some non-zero death and exposure counts at higher ages still. This is possible because the data at

different ages refer to different groups of cohorts. In this situation, we have chosen to terminate the life table for an almost-extinct cohort at the lowest age where $q_x = 1$.

Figure 13. Life table calculations for almost-extinct cohorts aged x^* to $x^* + 4$ in year t_n



Abridged life tables

For all life tables computed as part of this project, abridged tables are always extracted directly from complete tables. This process can be described in just two steps: 1) extract values of l_x, T_x, e_x for the abridged table directly from the complete table; and 2) compute ${}_nL_x = T_x - T_{x+n}$, ${}_nd_x = l_x - l_{x+n}$, and ${}_nq_x = \frac{{}_nd_x}{l_x}$. For such calculations, $x = 0$, 1, 5, 10, 15, … , 110. Of course, $n = 5$ except at both extremes of the age range. For the open interval, $n = \infty$ and $l_\infty = T_\infty = 0$.

Appendix A

Linear model for splitting 1x1 death counts

Two equations are given in the main text for use in splitting death counts into Lexis triangles based on deaths counts in 1x1 Lexis squares (see equations 4 and 7). These equations were derived from a multiple regression analysis, as summarized in Table A-1. In this appendix, we give more detail about the regression analysis, but without repeating formulas already presented in the main text.

The regression analysis was performed separately for men and women. The dependent variable was the proportion of deaths occurring in the lower triangle out of the total in a 1x1 Lexis square, or $\pi_d(x,t)$ (see equation 3). The analysis included data for ages 0-104 from Sweden (1901-1999), Japan (1950-1998), and France (1907-1997). During these time periods, death counts in Lexis triangles are available across the age range in the raw data with only minor exceptions.[23] A series of regression models was fitted by weighted least squares.

The weight associated with each observed value of $\pi_d(x,t)$ was defined as follows:

$$w(x,t) = \frac{D(x,t)}{\sum_x D(x,t)} \quad . \qquad (A1)$$

Thus, the total weight for a given country in a given year was one. Alternatively, we might have used $w(x,t) = D(x,t)$, motivated by the knowledge that the variability of $\pi_d(x,t)$ is inversely related to $D(x,t)$. However, such a choice also has the effect of giving much more weight to the most populous country, Japan, which then dominates the analysis. The weights used here accord an equal importance to

[23] The available French data are not classified by Lexis triangle above age 100 for some years (1934-35, 1947, 1949, 1954, and 1956-67). In these cases, the data used as an input to the regression analysis had already been split into triangles by some method (in the data file provided by Jacques Vallin and France Meslé). Ideally, such data should have been excluded from this regression analysis. However, given the small number of observations involved and their small weight in the total analysis, their exclusion would have had only a minor effect on the estimated coefficients, and then only for the age group 100-104. On the other hand, the computer programming was simplified by leaving these observations in the analysis, and so they were included.

each country-year included in the analysis, while at the same time giving more weight within each country-year to observations derived from larger numbers of deaths.

The variables included in the models were chosen after an extensive exploratory analysis. Model I includes only age effects, which reflect the changing level of $\pi_d(x,t)$ across the age range and explain around 70 percent of the variability in the dependent variable. The proportion of births associated with the lower triangle, $\pi_b(x,t)$ (see footnote 11 and equation 5), improves the fit further, as seen in Model II. The Spanish flu epidemic during the winter of 1918-1919 had the effect of increasing the proportions of lower-triangle deaths in 1918 (which includes more deaths from the second half of the calendar year) and of upper-triangle deaths in 1919 (for the opposite reason). Since this was a global epidemic, it seems reasonable to extrapolate the experience of Sweden and France (Japanese data begin later) onto the rest of the populations in the HMD.

For most age groups, the dependent variable, $\pi_d(x,t)$, has tended to increase over time, presumably in relation to changing levels and patterns of mortality. Partly as a matter of convenience, the infant mortality rate (IMR) was chosen to serve as a proxy variable for these sorts of temporal changes (Model IV). The IMR is convenient for this purpose because it can be estimated using only birth and infant death counts by calendar year. As explained in the main text, we used a simple method proposed by Pressat (1980) for computing the IMR. In this method, the denominator of the IMR is composed of two-thirds of the births from the current year plus one-third of the births from the previous year, since infant deaths in the current year are derived from both of these cohorts (although more were born in the current year than in the previous one).

The logarithmic transformation was used because it makes the relationship between the IMR and $\pi_d(x,t)$ more nearly linear across a broader range of observations. Even with this transformation, however, two other adjustments were needed to obtain a model that reflects well the patterns in the raw data. First, in Model VII (not shown here), we tested for possible interactions between $\log IMR$ and each age group. These interaction terms were statistically and practically significant for ages zero and one

only (Model V). Second, at age zero the relationship between $\log IMR$ and $\pi_d(x,t)$ seems to turn around at very low values of the IMR. Therefore, we added an interaction term (for this age only) between $\log IMR$ and a dummy variable to indicate when the IMR is below 0.01 (Model VI). This cut-off level was chosen to maximize the R-squared statistic. For both males and females, R-squared obtained a maximum value (with four decimal points of precision) for a cut-off value in the range of 0.009 to 0.011, approximately. Therefore, the value of 0.01 was used for both sexes.

The color graphs on the following pages show actual values of $\pi_d(x,t)$ at ages 0 and 80 for the three countries, along with predictions from the model. For each age, two graphs are shown, depicting the changes in $\pi_d(x,t)$ as a function of both time and $\log IMR$. In the graphs organized by time, we also show the predicted trend in $\pi_d(x,t)$ for Sweden prior to the observation period, since the infant mortality rate was available back to 1751 and a birth series back to 1749. As illustrated by these graphs, the average value of $\pi_d(x,t)$ stabilizes at high values of IMR (due to the logarithmic transformation), so there should be no problem with applying this model to historical periods. In the backwards extrapolation for age 80, the lack of a birth series prior to 1749 is immediately apparent. For earlier cohorts, it was necessary to assume a constant cohort size, resulting in a loss of variability in predicted $\pi_d(x,t)$.

Figure A-1. Proportion of male infant deaths in lower triangle

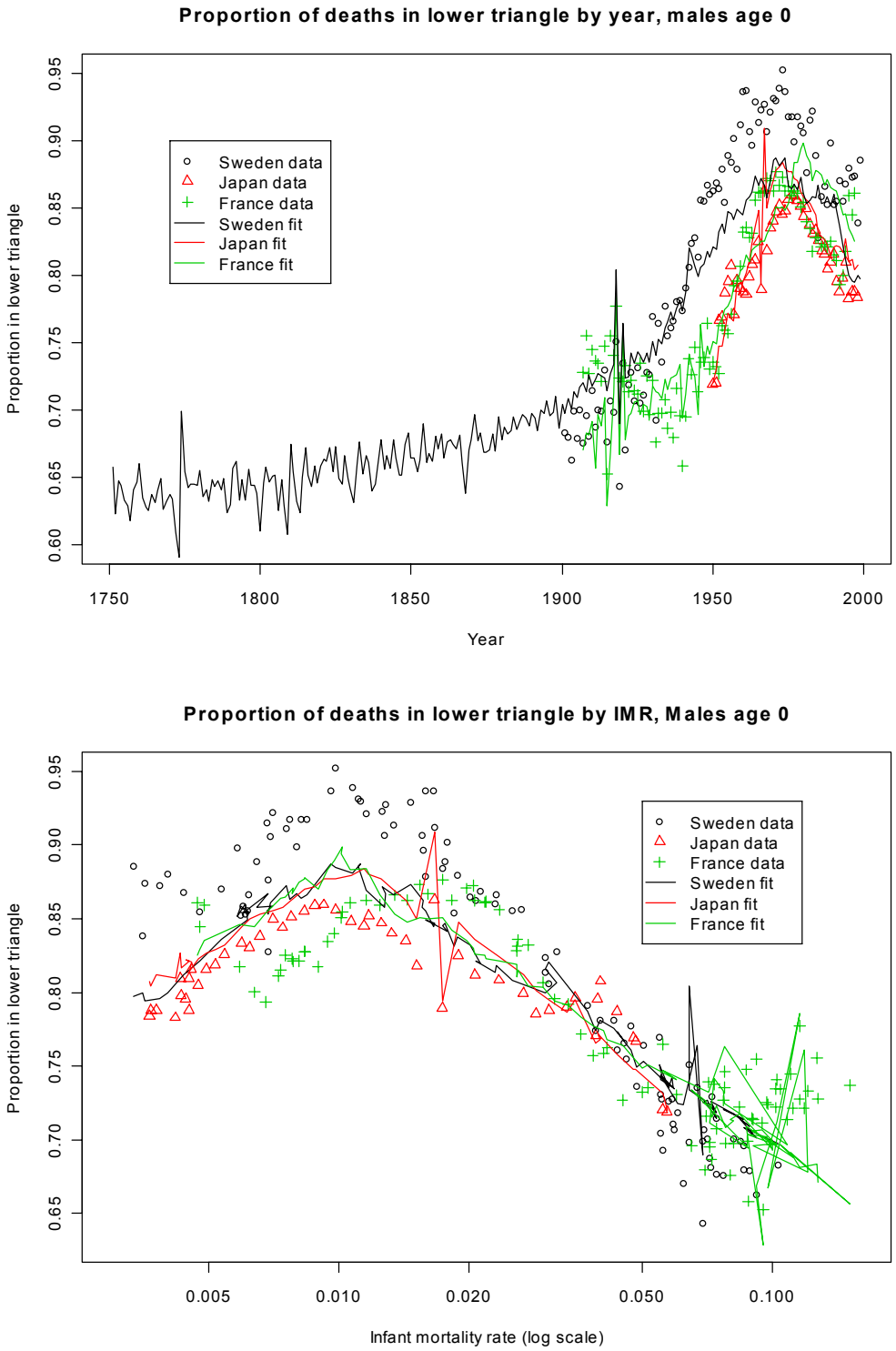


Figure A-2. Proportion of male age 80 deaths in lower triangle

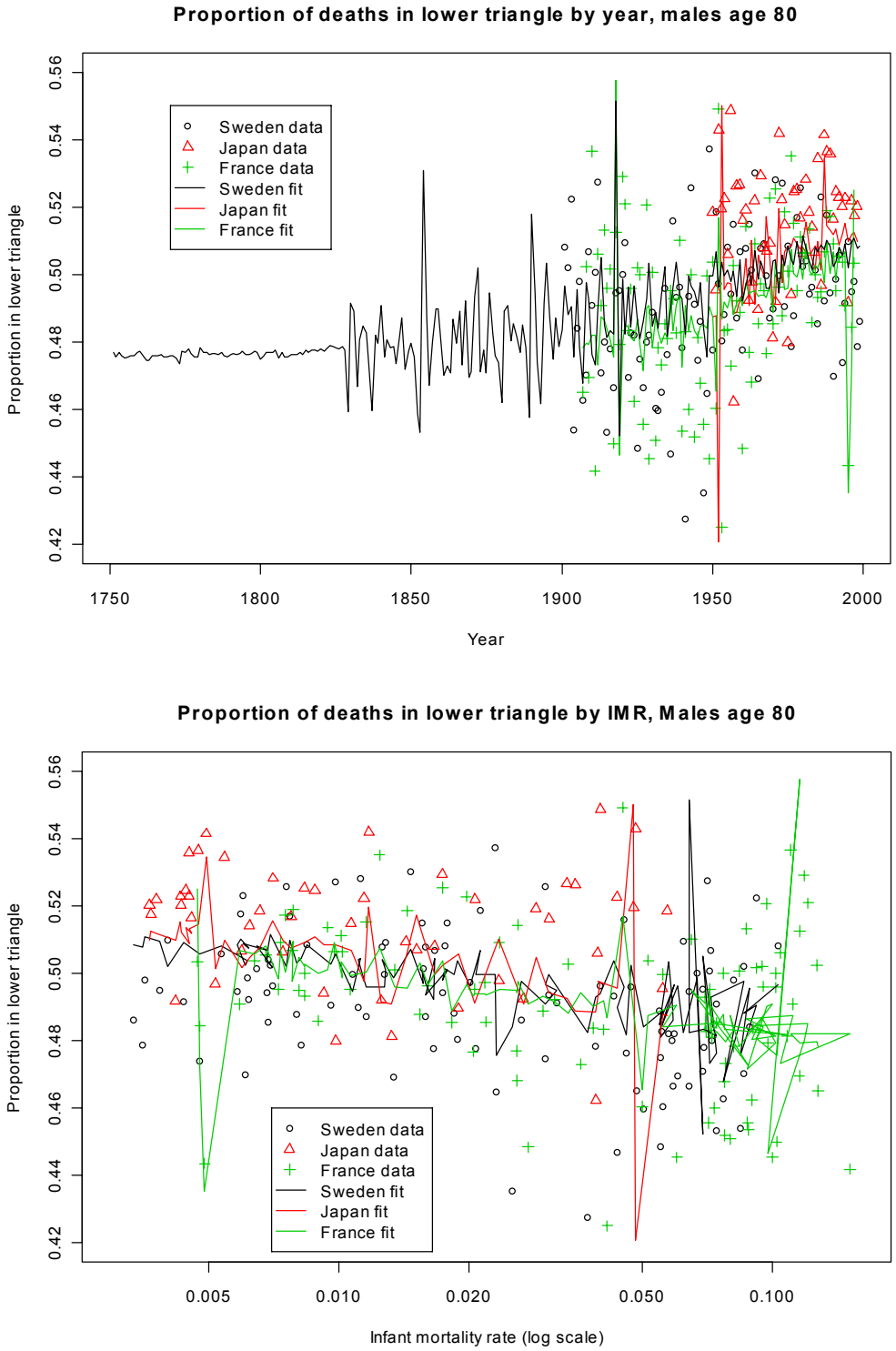


Table A-1. Linear models of the proportion of lower-triangle deaths [a]

a) Women (n = 25,017) [b]

	Null	I	II	III	IV	V	VI
Intercept	0.5104	0.5170	0.5172	0.5163	0.4712	0.4724	0.4710
Age groups [c]							
0		0.2285	0.2295	0.2299	0.2368	0.0697	0.0392
1		0.0462	0.0480	0.0479	0.0570	0.1351	0.1365
2-4		-0.0003	0.0011	0.0006	0.0084	0.0116	0.0130
5-9		-0.0101	-0.0092	-0.0100	-0.0029	0.0004	0.0018
10-14		-0.0254	-0.0246	-0.0258	-0.0186	-0.0154	-0.0140
15-19		-0.0242	-0.0235	-0.0249	-0.0181	-0.0149	-0.0135
20-24		-0.0162	-0.0159	-0.0171	-0.0108	-0.0074	-0.0061
25-29		-0.0141	-0.0137	-0.0151	-0.0093	-0.0059	-0.0046
30-34		-0.0127	-0.0126	-0.0137	-0.0090	-0.0055	-0.0041
35-39		-0.0153	-0.0152	-0.0157	-0.0123	-0.0086	-0.0072
40-44		-0.0142	-0.0142	-0.0142	-0.0123	-0.0084	-0.0070
45-49		-0.0131	-0.0134	-0.0131	-0.0126	-0.0085	-0.0071
50-54		-0.0136	-0.0142	-0.0138	-0.0140	-0.0098	-0.0084
55-59		-0.0140	-0.0145	-0.0140	-0.0148	-0.0105	-0.0091
60-64		-0.0180	-0.0185	-0.0179	-0.0191	-0.0148	-0.0134
65-69		-0.0215	-0.0221	-0.0215	-0.0233	-0.0189	-0.0175
70-74		-0.0233	-0.0240	-0.0234	-0.0260	-0.0215	-0.0201
75-79		-0.0251	-0.0258	-0.0251	-0.0291	-0.0244	-0.0230
80-84		-0.0235	-0.0240	-0.0233	-0.0294	-0.0245	-0.0231
85-89		-0.0165	-0.0173	-0.0165	-0.0253	-0.0201	-0.0187
90-94		-0.0066	-0.0073	-0.0065	-0.0181	-0.0125	-0.0112
95-99		0.0055	0.0047	0.0055	-0.0086	-0.0027	-0.0014
100-104		0.0274	0.0267	0.0275	0.0114	0.0176	0.0190
Birth proportion – 0.5 [d]			0.7255	0.7220	0.7357	0.7377	0.7372
Year = 1918				0.0887	0.1019	0.1023	0.1025
Year = 1919				-0.0379	-0.0243	-0.0239	-0.0237
log *IMR*					-0.0127	-0.0111	-0.0112
(log *IMR*) x (Age = 0)						-0.0571	-0.0688
(log *IMR*) x (Age = 1)						0.0268	0.0268
(log *IMR* – log 0.01) x (Age = 0) x (*IMR* < 0.01)							0.1526
R-squared [e]	0.0000	0.7113	0.7362	0.7558	0.7941	0.8170	0.8192

[a] All models were fit by weighted least squares, with weights equal to the number of deaths in a 1x1 Lexis square divided by the total deaths for that country in that year. Thus, the total weight for each country-year combination is one (see main text of Appendix A for discussion).
[b] All models were fit to data for ages 0-104 for Sweden (1901-1999), Japan (1950-1998), and France (1907-1997), after eliminating 78 observations with zero deaths in the 1x1 Lexis square.
[c] Since coefficients for age groups are constrained to sum to zero, there is no omitted category.
[d] The birth proportion equals the number of births in the younger cohort (born in t) divided by the total for the younger and older cohort (born in t-1). Data are centered about 0.5 (i.e., 0.5 is subtracted from the birth proportion for each observation).
[e] R-squared here is the proportion of weighted variance (about the weighted mean) explained by the model.

Table A-1. **Linear models of the proportion of lower-triangle deaths** [a]

b) Men (n = 24,872) [b]

	Null	I	II	III	IV	V	VI
Intercept	0.5176	0.5226	0.5226	0.5223	0.4831	0.4853	0.4838
Age groups [c]							
0		0.2294	0.2300	0.2304	0.2371	0.0555	0.0230
1		0.0370	0.0382	0.0385	0.0469	0.1234	0.1249
2-4		-0.0044	-0.0034	-0.0033	0.0035	0.0071	0.0086
5-9		-0.0088	-0.0080	-0.0080	-0.0021	0.0016	0.0031
10-14		-0.0200	-0.0193	-0.0194	-0.0139	-0.0101	-0.0086
15-19		-0.0274	-0.0269	-0.0274	-0.0230	-0.0190	-0.0175
20-24		-0.0056	-0.0057	-0.0069	-0.0019	0.0020	0.0035
25-29		-0.0014	-0.0006	-0.0019	0.0027	0.0066	0.0081
30-34		-0.0056	-0.0053	-0.0062	-0.0025	0.0016	0.0031
35-39		-0.0145	-0.0143	-0.0148	-0.0124	-0.0080	-0.0065
40-44		-0.0189	-0.0188	-0.0188	-0.0179	-0.0132	-0.0117
45-49		-0.0210	-0.0212	-0.0210	-0.0212	-0.0163	-0.0148
50-54		-0.0201	-0.0204	-0.0202	-0.0211	-0.0160	-0.0145
55-59		-0.0195	-0.0197	-0.0194	-0.0209	-0.0157	-0.0142
60-64		-0.0206	-0.0208	-0.0205	-0.0225	-0.0172	-0.0157
65-69		-0.0221	-0.0225	-0.0222	-0.0247	-0.0193	-0.0179
70-74		-0.0236	-0.0241	-0.0238	-0.0268	-0.0213	-0.0198
75-79		-0.0255	-0.0260	-0.0257	-0.0294	-0.0238	-0.0223
80-84		-0.0240	-0.0244	-0.0241	-0.0289	-0.0231	-0.0216
85-89		-0.0169	-0.0176	-0.0173	-0.0236	-0.0175	-0.0160
90-94		-0.0077	-0.0085	-0.0081	-0.0162	-0.0098	-0.0083
95-99		0.0063	0.0054	0.0057	-0.0043	0.0024	0.0039
100-104		0.0348	0.0339	0.0343	0.0229	0.0299	0.0313
Birth proportion – 0.5 [d]			0.6798	0.6778	0.6929	0.6992	0.6992
Year = 1918				0.0611	0.0725	0.0725	0.0728
Year = 1919				-0.0481	-0.0362	-0.0355	-0.0352
log *IMR*					-0.0108	-0.0088	-0.0088
(log *IMR*) x (Age = 0)						-0.0620	-0.0745
(log *IMR*) x (Age = 1)						0.0259	0.0259
(log *IMR* – log 0.01) x (Age = 0) x (*IMR* < 0.01)							0.1673
R-squared [e]	0.0000	0.6963	0.7163	0.7264	0.7492	0.7743	0.7768

See notes for Table A-1a.

Appendix B

Computational methods for fitting cubic splines

Splitting *n*x1 data into 1x1 format

Aggregated death counts are split into a 1x1 format using cubic splines fitted to the cumulative distribution of deaths within each calendar year. Let $Y(x) = \sum_{u=0}^{x-1} D_u$ be the cumulative number of deaths up to age *x*, and assume that $Y(x)$ is known for a limited collection of ages always including $x = 1$ and $x = 5$. Following McNeil *et al.* (1977), we fit a cubic spline to $Y(x)$ in the form of the following equation:

$$Y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \beta_1 (x - k_1)^3 I(x > k_1) + \cdots + \beta_n (x - k_n)^3 I(x > k_n) \ , \qquad (B1)$$

where $\alpha_0, \ldots, \alpha_3, \beta_1, \ldots, \beta_n$ are coefficients that must be estimated. The indicator function, $I(.)$, equals one if the logical statement within parentheses is true and zero if it is false. Along the *x*-axis, there are *n* "knots" denoted by k_i, $i = 1, \ldots, n$, and lower and upper boundaries denoted by *a* and *b*, respectively. In general, the knots are those values of *x* for which $Y(x)$ is known from the data, except for the lowest and highest such values.

We require that $k_1 = 1$, and k_n equals the lower limit of the open age interval. We always have $a = 0$ and $b = \omega$, where ω is set arbitrarily to the maximum of 105 or $k_n + 5$.[24] Thus, we know $n + 2$ values of $Y(x)$, for $x = 0, 1, \ldots, k_n$, and ω, but the above equation contains $n + 4$ unknown parameters. Therefore, two additional constraints are needed in order to compute the coefficients. Typical solutions usually involve constraining the slope of the function at the boundaries. At the upper boundary, for example, we constrain the slope to be zero. Thus, $Y'(\omega) = 0$. This choice is consistent with the usual tapering of the distribution of deaths at the oldest ages. However, a similar constraint at the lower

[24] Note that this choice makes no difference, since we do not use the fitted spline curve to split death counts in the open age interval anyway.

boundary would not be appropriate, since deaths are highly concentrated at age 0. Instead, we constrain the slope of the function at age 1 to equal one half the increment (in cumulative deaths) between ages 1 and 5. Thus, $Y'(1) = \frac{Y(5) - Y(1)}{2}$. Since $Y'(1) \approx D(1)$, this formula is based on the observation that about half of all deaths between ages 1 and 5 tend to occur during the second year of life (at all levels of mortality). The first derivative of $Y(x)$ is as follows:

$$Y'(x) = \alpha_1 + 2\alpha_2 x + 3\alpha_3 x^2 + 3\beta_1 (x - k_1)^2 I(x > k_1) + \cdots + 3\beta_n (x - k_n)^2 I(x > k_n) \quad . \text{(B2)}$$

Fitting the cubic spline function consists of solving a system of $n + 4$ linear equations. These equations can be written as follows:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & k_1 & k_1^2 & k_1^3 & 0 & 0 & \cdots & 0 & 0 \\ 1 & k_2 & k_2^2 & k_2^3 & (k_2 - k_1)^3 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & k_n & k_n^2 & k_n^3 & (k_n - k_1)^3 & (k_n - k_2)^3 & \cdots & (k_n - k_{n-1})^3 & 0 \\ 1 & \omega & \omega^2 & \omega^3 & (\omega - k_1)^3 & (\omega - k_2)^3 & \cdots & (\omega - k_{n-1})^3 & (\omega - k_n)^3 \\ 0 & 1 & 2 & 3 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 2\omega & 3\omega^2 & 3(\omega - k_1)^2 & 3(\omega - k_2)^2 & \cdots & 3(\omega - k_{n-1})^2 & 3(\omega - k_n)^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_3 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} 0 \\ Y(1) \\ \vdots \\ Y(k_n) \\ Y(\omega) \\ [Y(5) - Y(1)]/2 \\ 0 \end{pmatrix} \quad . \text{(B3)}$$

Writing this equation as $\mathbf{Ac} = \mathbf{d}$, the vector of coefficients can be found by computing $\mathbf{c} = \mathbf{A}^{-1}\mathbf{d}$. Once the coefficients are computed by this method, the estimated equation is used to find fitted values, $\hat{Y}(x)$, for $x = 0$, 1, 2, … , k_n. For all ages below the open age interval, deaths by single years of age are estimated by differencing:

$$\hat{D}_x = \hat{Y}(x+1) - \hat{Y}(x) \quad , \qquad \text{(B4)}$$

for $x = 0$, 1, 2, … , k_{n-1}.

The choice of constraints is very important. One drawback of the spline method is that the fitted curve may not be monotonically increasing over all ages. Since the curve depicts the cumulative deaths over age, a decreasing function between ages *x* and *x*+1 implies negative death counts at age *x*. We have tried to choose constraints that minimize the possibility of such an occurrence. Nevertheless, there seems

to be no reliable solution at the oldest ages, and the spline function often starts to decline within the open age group. For this reason, we use a different method for splitting deaths from an open age interval into finer categories. Fortunately, the constraint applied to the slope at age 1 seems to work in all cases, yielding a curve that is always monotonically increasing at younger ages.

Splitting period-cohort data covering multiple cohorts

With minor modifications, the method described above can be used to split period-cohort data covering multiple cohorts (usually in the shape of a parallelogram) into data for single-year birth cohorts. For example, suppose we know the values of $Y(1), Y(4.5), Y(9.5), \ldots, Y(99.5)$, where by definition $Y(x+\frac{1}{2}) = \sum_{j=0}^{x-1}[D_L(j) + D_U(j)] + D_L(x)$ is the cumulative number of deaths up to and including the lower triangle of age *x*. Then, we can fit a cubic spline with knots at those values. Because $Y(4.5)$ is known instead of $Y(5)$, we use a modified constraint: $Y'(1) = \frac{Y(4.5) - Y(1)}{1.83}$.[25]

From the fitted values, $\hat{Y}(x+\frac{1}{2})$, we obtain estimates of deaths for single-year cohorts by computing first differences. Then, following our usual practice (see footnote 10), the resulting period-cohort death counts are split 50/50 to obtain estimated death counts by Lexis triangle. These two steps can be summarized as follows:

$$\hat{D}_U(x) = \hat{D}_L(x+1) = \frac{1}{2}[\hat{Y}(x+\tfrac{3}{2}) - \hat{Y}(x+\tfrac{1}{2})] \text{ , for } x \geq 1 \text{ .} \qquad \text{(B5)}$$

Finally, we derive $\hat{D}_L(1) = \hat{Y}(1.5) - D(0)$.

[25] This modification derives from assuming that the deaths between exact ages 2 and 5 are uniformly distributed. Following our earlier logic, these deaths comprise half of all deaths between ages 1 and 5. If they are uniformly distributed, losing the upper triangle at age 4 means we are missing $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$ of all deaths between ages 1 and 5. Thus, we have:

$$Y'(1) = \frac{Y(5) - Y(1)}{2} = \frac{\frac{12}{11}[Y(4.5) - Y(1)]}{2} = \frac{Y(4.5) - Y(1)}{1.83} .$$

Appendix C

Method for splitting deaths in an open age interval

Our method for splitting deaths in an open age interval treats deaths above age $x^* - 20$ as though they come from a stationary population (where x^* is the lower boundary of the open age interval). Accumulating deaths backwards from the open age interval within a given calendar year, we divide by all deaths at ages $x^* - 20$ and older to get the *observed* cumulative proportion of deaths in that range that lie above age $x^* - i$:

$$S(x^* - i) = \frac{{}_{\infty}D_{x^*-i}}{{}_{\infty}D_{x^*-20}}, \text{ for } i = 0, 1, \ldots 20, \qquad (C1)$$

where ${}_{\infty}D_{x^*-i}$ is the number of deaths in the age interval from age $x^* - i$ and above (including the open age interval). This procedure yields a fictitious survival function (conditional on survival to age $x^* - 20$), corresponding to a kind of "synthetic extinct cohort" (i.e., an extinct cohort based on period death counts).

We fit the Kannisto model of old-age mortality (Thatcher et al., 1998) to this fictitious survival function. The Kannisto model has the following form:

$$\mu(x) = \frac{ae^{b(x-x_0)}}{1 + ae^{b(x-x_0)}}, \qquad (C2)$$

where $x_0 = x^* - 20$, and a and b are unknown parameters. The corresponding survival function is given by:

$$s(x) = \left(\frac{1 + a}{1 + ae^{b(x-x_0)}} \right)^{1/b}. \qquad (C3)$$

To summarize, we treat deaths within the calendar year as though they occurred in a stationary population with an age-specific pattern of mortality following the Kannisto model. Therefore, by

assumption, the predicted proportion of deaths in the population above age x is:

$$\hat{S}(x) = \hat{s}(x) \ , \ \text{for } x = x_0 \, , \ x_0 + 1 \, , \ldots \omega. \qquad (C4)$$

We fit the model in order to estimate parameters *a* and *b*.[26] Using equation (C3) and these estimates of $\hat{a}$ and $\hat{b}$, we predict:

$$d(x) = \hat{s}(x) - \hat{s}(x + 0.5) \ , \ \text{for } x = x^* \, , \ x^* + 0.5 \, , \ x^* + 1 \, , x^* + 1.5 \, , \ldots \omega. \qquad (C5)$$

Then, we apply these $d(x)$ to ${}_{\infty}D_{x^*}$ (and divide by the proportion surviving to age x^*) in order to derive the distribution of deaths by Lexis triangle within the open age interval beginning at age x^*:

$$\hat{D}_L(x) =_{\infty} D_{x^*} \cdot \frac{d(x)}{\hat{s}(x^*)} \ \text{ and } \ \hat{D}_U(x) =_{\infty} D_{x^*} \cdot \frac{d(x+0.5)}{\hat{s}(x^*)} \ , \text{ for } x = x^* \, , \ x^* + 1 \, , \ldots \omega. \qquad (C6)$$

If the estimated number of deaths in a given Lexis triangle is less than 0.25, we assume there are no deaths in that triangle or above. The estimated deaths at ages below that triangle (within the open age interval) are then adjusted proportionally (i.e., multiplied by a constant) so that their sum is equal to ${}_{\infty}D_{x^*}$. This entire procedure is applied separately to male and female death counts and then total death counts are obtained by summing.

Correction for Unusual Fluctuations in Deaths

In some cases, there may be an unusual fluctuation in death counts within the age range (from $x^* - 20$ to x^*) used to fit the Kannisto model. For example, a cohort in this age range may be particularly small relative to nearby cohorts because of some historical event (e.g., a war) and thus, experience fewer deaths. In such cases, we may want to exclude that outlier before fitting the model, or else we are likely to under-estimate the number of deaths at the start of the open age interval. In order to

[26] To fit the model, we minimize the squared differences between the logarithms of observed and predicted cumulative proportions:

$$\sum_{i=0}^{19} [\ln S(x^* - i) - \ln \hat{S}(x^* - i)]^2$$

stopping at i = 19 because $S(x^*–20)$ and $\hat{S}(x^*–20)$ equal 1.0 by definition.

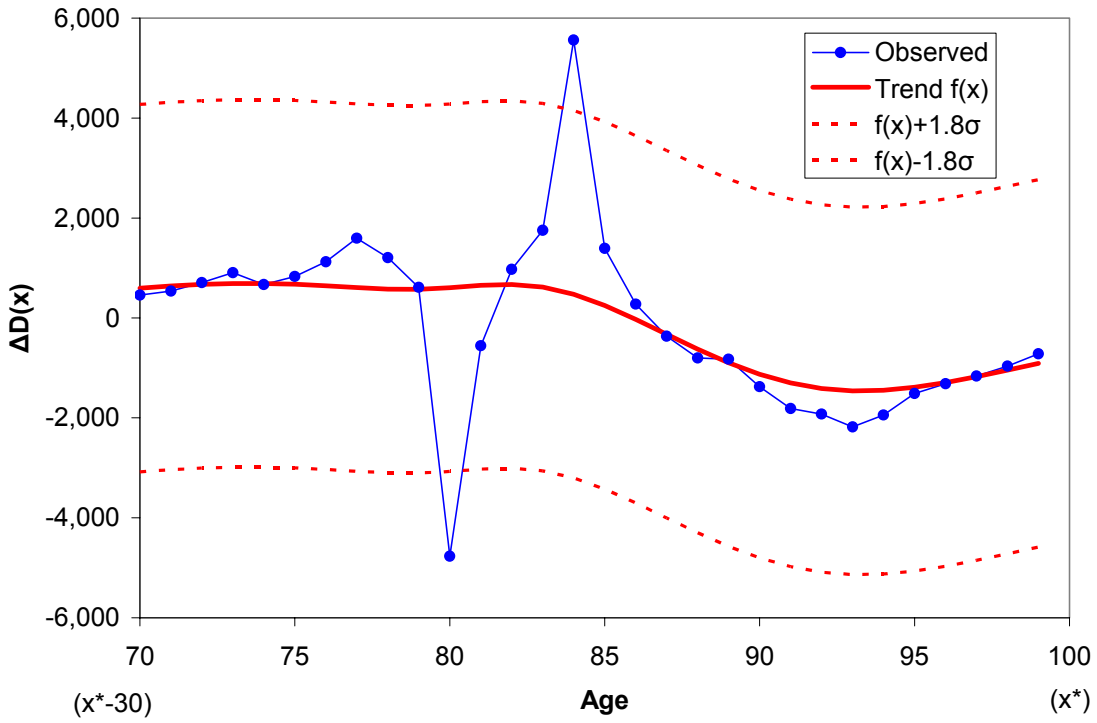
avoid this problem, we introduce a correction to the method described above. Prior to fitting the model, we first apply a procedure to identify any outliers, and if found, exclude the age range associated with the period of unusual fluctuations before fitting the model. The procedure for identifying outliers proceeds as follows:

1. We calculate first differences in deaths $\Delta D(x)$:

$$\Delta D(x) = D(x+1) - D(x) \ , \ \text{for } x = x^* - 30 \, , \, x^* - 29 \, , \, \ldots x^* - 2 \ . \qquad (C7)$$

Figure C-1 shows an example where $\Delta D(80)$ has a large negative value (i.e., there was a large decrease in deaths between ages 80 and 81) and $\Delta D(84)$ has a large positive value (i.e., there was a large increase in deaths between ages 84 and 85).

Figure C-1. First differences in deaths, West German females, 1999



2. We identify the trend in $\Delta D(x)$ by fitting a cubic smoothing spline that minimizes the following function:

$$p \sum_{x=x^*-30}^{x^*-2} \left(\Delta D(x) - f(x)\right)^2 + (1-p) \int_{x^*-30}^{x^*-2} \left(f''(x)\right)^2 dx \ , \qquad (C8)$$

where p is the smoothing parameter and $f(x)$ is a standard cubic spline function (e.g., see McNeil et al., 1977 and equation B1). As p tends toward 0, the result approaches the least-squares linear fit (i.e., a straight line), whereas for $p = 1$, the result is the natural cubic spline. As p varies from 0 to 1, the result moves from one extreme to the other. We use $p = 0.0005$ because it provides a suitable compromise between a trend that would identify too many fluctuations as outliers (e.g., linear fit) and one that requires the trend to pass through each of the observed data points (i.e., no outliers are identified). Figure C-1 shows the trend $f(x)$ plotted against observed values of $\Delta D(x)$.

3. We define $\Delta D(x)$ to be an outlier if:

$$\left|\Delta D(x) - f(x)\right| > 1.8\sigma \ , \qquad (C9)$$

where σ represents the standard deviation of the difference between $\Delta D(x)$ and $f(x)$. Figure C-1 shows $f(x) \pm 1.8\sigma$. Any $\Delta D(x)$ that falls outside the range of plus or minus 1.8 standard deviations is defined as an outlier (e.g., $\Delta D(80)$ and $\Delta D(84)$ in this example).[27]

4. The age range associated with each period of unusual fluctuations will be identified by two outliers on $\Delta D(x)$: one at the age where the fluctuation begins $x_{\min}$, and one at the age where it ends $x_{\max}$. We then substitute the interpolated values, based on a cubic smoothing spline similar to equation C8), for all observations from ages $x_{\min}$ to $x_{\max}$ before fitting the model.[28] So, in

[27] We tested this procedure using data for France, Sweden, and Czech Republic. Based on visual inspection of the data, we identified observations that appeared to be outliers. Then, we set the parameters ($p = 0.0005$ and $\pm 1.8\sigma$) for the procedure such that we detected all "real" outliers (based on our subjective evaluation), but minimized the number of false positives. Nonetheless, in most cases, treating "false positives" as outliers did not change the mortality estimates, whereas excluding "real" outliers had a substantial effect on the estimates.

[28] All outliers identified by the procedure described here are also checked by the country specialist to ensure that we are not treating observations as outliers that should not be treated as such (i.e., we do not exclude detected outliers if doing so produces implausible results).

this example, we substitute the *estimated* proportion of survivors for the *observed* proportion of survivors $S(80)$, $S(81)$,… $S(84)$ (see equation C1) before fitting the model, using $\widetilde{D}(81), \widetilde{D}(82),...\widetilde{D}(85)$ instead of $D(81), D(82),...D(85)$, where $\widetilde{D}(x)$ is derived by minimizing the following function:

$$p \sum_{x=x^*-30}^{x^*-1} \left(D(x) - f(x)\right)^2 + (1-p) \int_{x^*-30}^{x^*-1} \left(f''(x)\right)^2 dx \,. \qquad (C10)$$

Correction for Cohort Size

Another potential problem with the basic method for splitting the open age interval is that sometimes deaths for a particular cohort *within the open age interval* may be under- or over-estimated by the model as a result of variations in cohort size. For example, a cohort may be unusually small (e.g., those born in time of war) relative to other cohorts in that same period; if that cohort falls within the open age interval, then the model may over-estimate deaths for that cohort. To address this problem, we make a final correction for cohort size after splitting deaths in the open age interval as described above. This correction essentially redistributes deaths to take account of the size of each respective cohort relative to other nearby cohorts within the open age interval.

For each cohort born in year $t-x$, where x is the age at last-birthday on December 31st of year t for $x \geq x^*$, we perform the following steps:

1. We compare deaths for that cohort in the year in which they would attain age x^*-1 relative to the average of deaths at the same age among nearby cohorts. For example, for $x = x^* + 3$ (the cohort represented by the period-cohort parallelogram in yellow in Figure C-2), we compute the ratio of deaths for that same cohort in year $t-4$ (in the uppermost red parallelogram) to the average of deaths at the same age for the two previous cohorts and two later cohorts (shown in blue). If all five cohorts had the same number of deaths in the years they would attain age x^*-1, then the resulting ratio would be 1.0.

2. We calculate similar ratios for ages $x^* - 5$ through $x^* - 2$ (as depicted in Figure C-2).
3. Then, we take the average of the ratios across these five age years ($x^* - 5$ to $x^* - 1$) as an estimate of the size of this cohort relative to nearby cohorts.
4. Next, we multiply the original estimate of deaths in the open age interval (derived using methods described in previous sections) by this adjustment ratio. Following the earlier example, the adjustment ratio based on the deaths shown in red in Figure C-2 is applied to the original estimates of $D_U(x^* + 2, t)$ and $D_L(x^* + 3, t)$ shown in yellow.
5. Finally, we make one last minor adjustment to ensure that the estimates add up to the original sum in the open age interval.

Of course, this procedure requires that the data are available to make these calculations. For the example shown in Figure C-2), we could not complete the calculations described above if the data series began in year $t - 4$. Therefore, given the estimate of $D_U(x-1,t)$ where $x - 1 \geq x^*$, or $D_L(x,t)$ where $x \geq x^*$, we make no correction for cohort size if $t_{\min} \geq t - (x - x^*)$, where $t_{\min}$ is the first year of the data series. Furthermore, if the original raw death counts for the period used to make these calculations are not available by (at least) one-year age groups, we make no correction for cohort size.

Given that $t_{\min} < t - (x - x^*)$, following the five steps listed above, we adjust the original estimates of $D_U(x-1,t)$ and $D_L(x,t)$ as follows in order to correct for fluctuations in cohort size:

$$\hat{D}_U(x-1,t) = r(x) \cdot D_U(x-1,t) \ , \quad \text{for } x = x^* + 1, x^* + 2, ...\omega, \quad \text{and} \qquad (C11)$$

$$\hat{D}_L(x,t) = r(x) \cdot D_L(x,t) \ , \quad \text{for } x = x^*, x^* + 1, x^* + 2, ...\omega \ , \qquad (C12)$$

where

$$r(x) = \frac{1}{n} \sum_{j=1}^{n} r(x)_j \ , \qquad (C13)$$

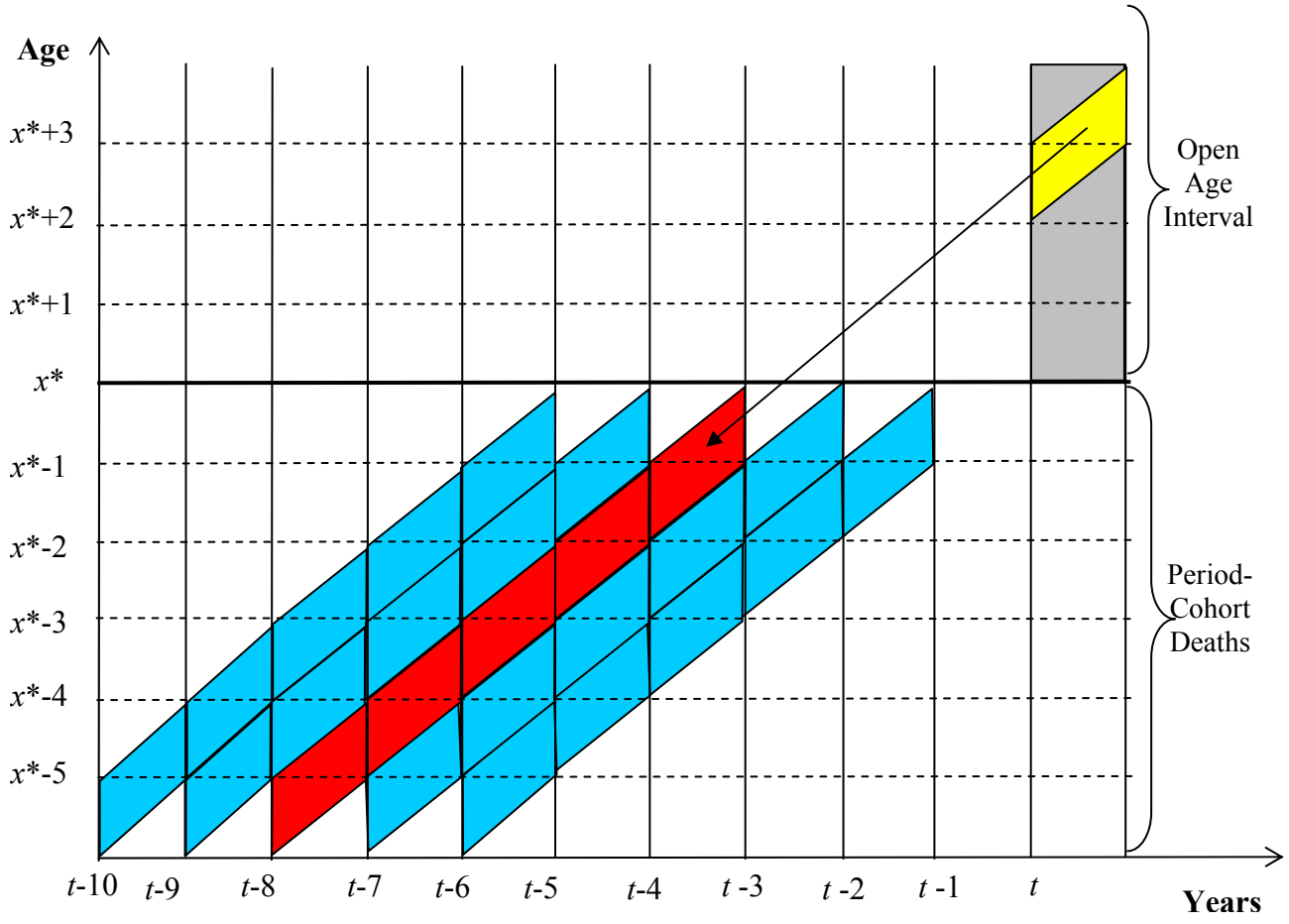
$$n = \min[(t - t_{\min} - (x - x^*)), 5] \ , \qquad (C14)$$

$$r(x)_j = \frac{D_U[x^* - j - 1, t - j - (x - x^*)] + D_L[x^* - j, t - j - (x - x^*)]}{\sum_{i=x-m}^{x+l} (D_U[x^* - j - 1, t - j - (i - x^*)] + D_L[x^* - j, t - j - (i - x^*)]) / (l + m + 1)} \quad , \text{(C15)}$$

$$l = \min[(t - t_{\min} - j - (x - x^*)), 2] \ , \qquad \text{(C 16)}$$

$$\text{and} \quad m = \min[(x - x^*), 2] \ . \qquad \text{(C17)}$$

Figure C-2. Example depicting the procedure to correct for cohort size



Then, we calculate the final estimates $\widetilde{D}_U(x-1,t)$ and $\widetilde{D}_L(x,t)$ using the following adjustment to ensure that they sum to the original death count in the open age interval:

$$\widetilde{D}_U(x-1,t) = \hat{D}_U(x-1,t) \cdot \frac{{}_{\infty}D_{x^*}(t)}{\sum_{i=x^*}^{\omega}[\hat{D}_L(i,t) + \hat{D}_U(i,t)]} \text{ and} \qquad \text{(C18)}$$

$$\widetilde{D}_L(x,t) = \hat{D}_L(x,t) \cdot \frac{{}_{\infty}D_{x^*}(t)}{\sum_{i=x^*}^{\omega}[\hat{D}_L(i,t) + \hat{D}_U(i,t)]} \,. \qquad \text{(C19)}$$

Appendix D

Adjustments for changes in population coverage

Some of the countries included in the HMD have experienced changes in their territorial boundaries. These changes must be taken into account when computing death rates and life tables (unless the changes are very small in relation to population size). In general, death counts must always refer to the same territory as the exposure-to-risk when calculating death rates. Likewise, when using birth counts as a measure of relative cohort size (in splitting 1x1 death counts into Lexis triangles), the birth series must refer to the same territory. With these principles in mind, we make some special calculations for countries with changing territories during the time period covered by the HMD.

Before describing these calculations, let us consider the format of the relevant data. Although territorial changes may occur at any time during a calendar year, administrative data (i.e., birth and death counts) typically reflect such changes on January 1st. Therefore, throughout this discussion we assume that the territorial change occurs on January 1st and that birth and death counts within an individual calendar year always refer to an unchanging territory.

Birth counts used in splitting 1x1 deaths

As described earlier, we use a birth series as a measure of relative cohort size as part of our method for splitting 1x1 death data into triangles. Suppose that this birth series is based on a changing territory. Then, the size of two successive cohorts may appear to change merely as a result of past territorial changes. If we use such a series in our calculations, we will introduce "artificial cohort effects" into our estimated death counts. Therefore, we adjust the birth series so that it refers to the same territory.

Suppose that a territorial change occurs on January 1st of year t. Define $R_b(t) = \frac{B^+}{B^-}$ to be the ratio of births in the new territory, B^+, to births in the old territory, B^-, either in year t or t-1 (see below). Therefore, in order to calculate $R_b(t)$ for the period covered by the birth series, we need birth data for the territory gained or lost during each territorial change. If territory was gained in year t, then we base the

calculation on the number of births in year t. B^- is calculated as the total births minus the births in the territory added, while B^+ is simply the total births in year t. On the other hand, if territory was lost in year t, then we base the calculation on the number of births in year $t-1$. B^- is simply the total births (including the area lost) in year $t-1$, while B^+ is calculated as the total births in year $t-1$ minus the births for the territory lost.[29] Thus, in either case, the data for both the numerator and the denominator come from the same year (either *t* or *t*-1). For example, in 1954, the territory of Trieste was added to the Italian territory. Therefore, we calculate $R_b(1954)$ as the ratio of the births (in 1954) for the entire territory (including Trieste), B^+, to the births (in 1954) for the territory excluding Trieste, B^-, resulting in $R_b(1954)=1.003$. If there was no territorial change in year t, then $R_b(t)=1.0$ by default.

The formulas shown in equations 5 and 6 (p. 13) are then modified as follows:[30]

$$\pi_b(x,t)=\frac{B(t-x)}{B(t-x)+B(t-x-1)\cdot R_b(t-x)}\ , \qquad (D1)$$

and

$$IMR(t)=\frac{D(0,t)}{\frac{1}{3}B(t-1)\cdot R_b(t)+\frac{2}{3}B(t)}\ . \qquad (D2)$$

Note that if there is no territorial change in year $t-x$, then $R_b(t-x)=1.0$ by default and thus, drops out of equation D1 leaving it exactly as shown in equation 5. Similarly, if there is no territorial change in year t, then $R_b(t)$ drops out of equation D2 and the result is identical to equation 6.

Ideally, we would like to have a birth series back to the earliest cohort for which we have death data (≈ 100 years prior to the earliest calendar year of deaths). In that case, we need $R_b(t)$ factors back to

[29] In some cases, the necessary birth data may not be available. In such cases, we simply use the population adjustment factor at age 0 (see p. 65): $R_b(t)=V(0,t)$.

[30] The birth ratio calculated in Equation D1 is used in order to split deaths in year *t*, based on data from the years in which the respective cohorts were born (i.e., years *t*-*x* and *t*-*x*-1). Yet, if a territorial change occurred between years *t*-*x* and year *t*, then the territory covered at the time of death, year *t*, is not the same as the territory covered at the time these cohorts were born. Equation D1 implicitly assumes that the birth ratio for the two cohorts is the same for the territory covered in year *t*-*x* as in the territory covered in year *t*.

the beginning of the birth series (if there were territorial changes). For example, suppose we have death data for 1900-2000 in country X and there was a territorial change in 1850. Individuals who died at age 50 in 1900 were born in either 1849 or 1850. In order to calculate the birth proportion, $\pi_b(50,1900)$, shown in equation (D1), we need $R_b(1850)$. Nonetheless, in many cases we will not have birth data prior to the earliest death data, in which case we set $\pi_b(x,t) = 0.5$ by default and assume $B(t-1) = B(t)$ in order to calculate $IMR(t)$ (see p. 13).

Extinct cohort methods

When we estimate population sizes using extinct cohort methods, we apply a different form of adjustment for territorial changes. Suppose that we are estimating $P(x)$ by this method, and that some territorial change occurs at time *t*. Define $V(x,t) = \frac{P^+(x)}{P^-(x)}$ to be the ratio of the population size at age *x* just after this change (i.e., on January 1st of year *t*) to the comparable value just before the change (i.e., on December 31st of year *t*-1).

For the country in question, we require population counts by age (and sex if available) for the territory that is gained or lost during the territorial change as well as for the entire country. Preferably, we use population estimates near the time of the territorial change, but sometimes we may only have data from a census at time *t** (close to the territorial change at time *t*). If a territory was added, then we must use data from the subsequent census, whereas if territory was lost, then we use data from the census prior to the territorial change.

In some cases, the available data may be aggregated into age groups, in which cases the *V(x,t)* factor is calculated for the age group and then applied to each single year of age within that age group. In fact, such data may be preferable because random variations across age are smoothed. Therefore, even if data by single year of age are available, we may still calculate the *V(x,t)* factors by five-year age groups. In any case, at very high ages (e.g., age 90 and older), the *V(x,t)* factor is calculated using aggregate data even if data by single year of age are used at younger ages. Aggregating across an open age interval at

very old ages is necessary because the population counts by single year of age can become very small at high ages, resulting in *V(x,t)* factors that are very erratic (including even zero or undefined values). If data are not available by age, then we must use *V(t)* calculated from the total population of all ages.

If there are no further territorial changes during the life of this cohort, then we estimate

$$P^{+}(x,t) = \sum_{i=0}^{\infty} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right], \qquad (D3)$$

$$\text{and} \quad P^{-}(x,t) = \frac{P^{+}(x,t)}{V(x,t)}. \qquad (D4)$$

Now, suppose there is no territorial change at time *t*, but rather at time t_1, where $t_1 > t$. Define $N_1 = t_1 - t$ (i.e., the time until the territorial change). We estimate $P(x,t)$ as follows:

$$\begin{aligned} P(x,t) = & \sum_{i=0}^{N_1-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \\ & + \frac{1}{V(x+N_1,t_1)} \sum_{i=N_1}^{\infty} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \end{aligned} . \qquad (D5)$$

If there is also a territorial change at time *t*, then this formula gives the value for $P^{+}(x,t)$.

For cohorts who live through more than one territorial change (at older ages), the above formula requires a slight modification. For example, assume that *s* territorial changes occur at times t_1, …, t_s, where $t_s > \cdots > t_1 > t$. Define $N_1 = t_1 - t$, …, $N_s = t_s - t$. Then, we estimate $P(x,t)$ as follows:

$$\begin{aligned} P(x,t) = & \sum_{i=0}^{N_1-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \\ & + \frac{1}{V(x+N_1,t_1)} \sum_{i=N_1}^{N_2-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \\ & + \frac{1}{V(x+N_1,t_1)V(x+N_2,t_2)} \sum_{i=N_2}^{N_3-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \\ & + \cdots \\ & + \frac{1}{V(x+N_1,t_1)\cdots V(x+N_{s-1},t_{s-1})} \sum_{i=N_{s-1}}^{N_s-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \\ & + \frac{1}{V(x+N_1,t_1)\cdots V(x+N_s,t_s)} \sum_{i=N_s}^{\infty} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \end{aligned} . \qquad (D6)$$

Intercensal survival methods

When we estimate population sizes using intercensal survival methods, we apply a similar adjustment for territorial changes. Referring to Figures 4a-c, suppose that a change in the territorial coverage of vital statistics occurs on January 1st of year $t_1 = t + N_1$, where $N_1 \le N$ and N equals the number of complete calendar years between the two censuses. Given $V(x,t)$, defined as above, the main formulas for existing cohorts (Figure 4a) would be modified as follows:

$$\hat{C}_2 = \left\{ C_1 - (D_a + D_b) - \sum_{i=0}^{N_1-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \right\} \cdot V(x+N_1,t_1) - \sum_{i=N_1}^{N-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] - (D_c + D_d) \quad , \qquad (D7)$$

and

$$P(x+n,t+n) = \begin{cases} C_1 - (D_a + D_b) - \sum_{i=0}^{n-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] & \text{for } 0 \le n < N_1 \\ \quad + \dfrac{1 - f_1 + n}{N + 1 - f_1 + f_2} \cdot \dfrac{\Delta_x}{V(x+N_1,t_1)} & \\ \left\{ \begin{array}{l} C_1 - (D_a + D_b) \\ - \sum_{i=0}^{N_1-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] \end{array} \right\} \cdot V(x+N_1,t_1) & \\ \quad - \sum_{i=N_1}^{n-1} \left[D_U(x+i,t+i) + D_L(x+i+1,t+i) \right] & \text{for } N_1 \le n \le N \\ \quad + \dfrac{1 - f_1 + n}{N + 1 - f_1 + f_2} \Delta_x & \end{cases} . \qquad (D8)$$

Note that the latter equation provides two different estimates of $P(x+N_1,t_1)$, corresponding to the territory covered by the statistical system just before and after January 1st of year $t_1 = t + N_1$. Thus, $P^-(x+N_1,t_1)$ comes from the top part of the formula and $P^+(x+N_1,t_1)$ from the bottom part, and it is easy to confirm that $P^+(x+N_1,t_1) = P^-(x+N_1,t_1) \cdot V(x+N_1,t_1)$.

Similarly, for the infant cohort (Figure 4b), the modified formulas are as follows:

$$\hat{C}_2 = \left\{ C_1 - D_a - \sum_{i=0}^{N_1-1} \left[D_U(i,t+i) + D_L(i+1,t+i) \right] \right\} \cdot V(N_1,t_1) - \sum_{i=N_1}^{N-1} \left[D_U(i,t+i) + D_L(i+1,t+i) \right] - (D_c + D_d) \quad , \qquad (D9)$$

and

$$P(n,t+n)=\begin{cases} C_1-D_a-\sum_{i=0}^{n-1}\left[D_U(i,t+i)+D_L(i+1,t+i)\right] & \text{for } 0\le n<N_1 \\ \quad+\dfrac{\frac{1}{2}\left(1-f_1^2\right)+n}{N+\frac{1}{2}\left(1-f_1^2\right)+f_2}\cdot\dfrac{\Delta_0}{V(N_1,t_1)} & \\ \left\{C_1-D_a-\sum_{i=0}^{N_1-1}\left[D_U(i,t+i)+D_L(i+1,t+i)\right]\right\}\cdot V(N_1,t_1) & \text{for } N_1\le n\le N \\ \quad-\sum_{i=N_1}^{n-1}\left[D_U(i,t+i)+D_L(i+1,t+i)\right] & \\ \quad+\dfrac{\frac{1}{2}\left(1-f_1^2\right)+n}{N+\frac{1}{2}\left(1-f_1^2\right)+f_2}\Delta_0 & \end{cases} . \quad (D10)$$

Finally, we consider the case of new cohorts born during the intercensal period. For those born after the territorial change (i.e., in calendar year $t+N_1$ or later), the standard formulas can be used. For those born before the territorial change, however, modified formulas are needed. Consider the cohort born in year $t+j$ where $0\le j<N_1$. As before, define $K=N-j-1$ (the age of the cohort on January 1st before the second census). Also define $K_1=N_1-j-1$ (the age of the cohort on January 1st of year $t+N_1$). Then, the modified formulas are as follows:

$$\hat{C}_2=\begin{cases}\left\{B_{t+j}-D_L(0,t+j)-\sum_{i=1}^{K_1}\left[D_U(i-1,t+j+i)+D_L(i,t+j+i)\right]\right\}\cdot V(K_1,t_1) \\ \quad-\sum_{i=K_1+1}^{K}\left[D_U(i-1,t+j+i)+D_L(i,t+j+i)\right]-(D_c+D_d)\end{cases}, \quad (D11)$$

and

$$P(k,t+j+k+1)=\begin{cases} B_{t+j}-D_L(0,t+j) & \text{for } 0\le k<K_1 \\ \quad-\sum_{i=1}^{k}\left[D_U(i-1,t+j+i)+D_L(i,t+j+i)\right]+\dfrac{2k+1}{2K+1+2f_2}\cdot\dfrac{\Delta'_{t+j}}{V(K_1,t_1)} & \\ \left\{B_{t+j}-D_L(0,t+j)-\sum_{i=1}^{K_1}\left[D_U(i-1,t+j+i)+D_L(i,t+j+i)\right]\right\}\cdot V(K_1,t_1) & \text{for } K_1\le k\le K \\ \quad-\sum_{i=K_1+1}^{k}\left[D_U(i-1,t+j+i)+D_L(i,t+j+i)\right]+\dfrac{2k+1}{2K+1+2f_2}\Delta'_{t+j} & \end{cases} . \quad (D12)$$

Sometimes more than one territorial change occurs during an intercensal interval. In this situation, the formulas are only slightly more complicated. Suppose that s territorial changes occur at times $t_1 = t + N_1$, …, $t_s = t + N_s$. The formulas above are for the case where $s = 1$. If $s = 2$, the formulas for existing cohorts would be as follows:

$$\begin{aligned}\hat{C}_2 = &\left\{\begin{array}{l} C_1 - (D_a + D_b) \\ -\sum_{i=0}^{N_1-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] \end{array}\right\} \cdot V(x+N_1,t_1)\cdot V(x+N_2,t_2) \\ &-\sum_{i=N_1}^{N_2-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right]\cdot V(x+N_2,t_2) \\ &-\sum_{i=N_2}^{N-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] - (D_c + D_d)\end{aligned} \qquad \text{(D13)}$$

and

$$P(x+n,t+n) = \begin{cases} C_1 - (D_a + D_b) - \sum_{i=0}^{n-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] & \text{for } 0 \le n < N_1 \\ \quad + \dfrac{1 - f_1 + n}{N + 1 - f_1 + f_2} \cdot \dfrac{\Delta_x}{V(x+N_1)\cdot V(x+N_2)} & \\ \left\{\begin{array}{l} C_1 - (D_a + D_b) \\ -\sum_{i=0}^{N_1-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] \end{array}\right\} \cdot V(x+N_1) & \text{for } N_1 \le n < N_2 \\ \quad -\sum_{i=N_1}^{n-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] & \\ \quad + \dfrac{1 - f_1 + n}{N + 1 - f_1 + f_2} \cdot \dfrac{\Delta_x}{V(x+N_2)} & \\ \left\{\begin{array}{l} C_1 - (D_a + D_b) \\ -\sum_{i=0}^{N_1-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] \end{array}\right\} \cdot V(x+N_1)\cdot V(x+N_2) & \text{for } N_2 \le n \le N \\ \quad -\left\{\sum_{i=N_1}^{N_2-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right]\right\} \cdot V(x+N_2) & \\ \quad -\sum_{i=N_2}^{n-1}\left[D_U(x+i,t+i) + D_L(x+i+1,t+i)\right] & \\ \quad + \dfrac{1 - f_1 + n}{N + 1 - f_1 + f_2}\Delta_x & \end{cases} \qquad \text{(D14)}$$

These formulas can be adapted as well to cases where $s > 2$. Formulas for the infant cohort and new cohorts receive similar modifications to adjust for territorial changes.

Linear interpolation

When we use linear interpolation to calculate population size on January 1st (for example, when we have reliable population estimates referring to July 1st of each adjacent year), it is important to ensure that both populations refer to the same territory. When a territorial change occurs on January 1st of year *t,* we multiply the population estimate for the given date in year *t*-1 by $V(x,t)$ before performing linear interpolation. In this way, the population in year *t*-1 is adjusted to reflect the same territory as on January 1st of year *t.*

Period death rates around the time of a territorial change

In the event of a territorial change, the formula for the exposure-to-risk (shown in equation 49), which is used to calculate the period death rate, requires a minor modification. Suppose there is a territorial change on January 1st of year *t*+1. The exposure-to-risk in the age interval $[x, x+1)$ during calendar year *t* would be calculated as follows:

$$E_{xt}^{p} = \frac{1}{2}\left[P(x,t) + P^{-}(x,t+1)\right] + \frac{1}{6}\left[D_L(x,t) - D_U(x,t)\right], \qquad \text{(D15)}$$

where $P^{-}(x,t+1)$ is the population at age *x* just before the territorial change.

Cohort mortality estimates around the time of a territorial change

Territorial changes present a special problem for cohort life table calculations. Fortunately, the solution is fairly simple. Let D_L and D_U be the cohort death counts at some age just before and after a territorial change (assumed to occur on January 1st). Also, let P^{-} and P^{+} be population estimates (at the same age) just before and after this change. Define $D_U^{*} = \frac{P^{-}}{P^{+}} D_U$, which represents the upper-triangle deaths that would have occurred in the original territory. Compute death rates and life table quantities as follows:

$$M_x = \frac{D_L + D_U^{*}}{P^{-} + \frac{1}{3}\left(D_L - D_U^{*}\right)}, \; a_x = \frac{\frac{1}{3}D_L + \frac{2}{3}D_U^{*}}{D_L + D_U^{*}}, \text{ and } q_x = \frac{D_L + D_U^{*}}{P^{-} + D_L}. \qquad \text{(D16)}$$

These three quantities are mutually consistent, like other cohort quantities. Furthermore, equivalent values are obtained if instead we define $D_L^* = \frac{P^+}{P^-} D_L$, which represents the lower-triangle deaths that would have occurred in the new territory, and compute as follows:

$$M_x = \frac{D_L^* + D_U}{P^+ + \frac{1}{3}\left(D_L^* - D_U\right)},\ a_x = \frac{\frac{1}{3}D_L^* + \frac{2}{3}D_U}{D_L^* + D_U}, \text{ and } q_x = \frac{D_L^* + D_U}{P^+ + D_L^*} \ . \qquad (D17)$$

Thus, the result is the same whether this adjustment is made to the lower or the upper triangle. The method is also consistent with an estimate of q_x based on the French method of partial quotients, since

$$q_x = 1 - \frac{P^-}{P^- + D_L}\frac{P^+ - D_U}{P^+} \ . \qquad (D18)$$

Other changes in population coverage

Sometimes there may be changes in the coverage of demographic data that are not due to territorial changes, but can be treated as such in order to make the appropriate adjustments to the formulas. For example, in some countries the vital statistics collection system changed from covering the *de facto* population to covering the *de jure* population at some time *t*. In order to account for this change in the birth and death count data, we treat it as a territorial change and calculate $V(x,t)$ factors based on the *de jure* population and the *de facto* population at the time of the change in population coverage. We then use $V(0,t)$ as an estimate of $R_b(t)$.

Appendix E

Computing death rates and probabilities of death

The purpose of this appendix is to describe and justify the methods used for computing death rates and probabilities of death in the HMD. We consider the case where death counts are available by Lexis triangle and population estimates are available by single years of age for individual calendar years (see Appendices A and B for the methods used to split 1x1 or 5x1 data, if needed). A key assumption is that deaths are distributed uniformly within Lexis triangles, and so we begin by exploring the implications of this assumption. We then show how the main formulas used in computing cohort and period death rates and probabilities of death can be derived making use of this assumption.

Uniform distribution of deaths

We typically assume that deaths are distributed uniformly within individual Lexis triangles. The main results following from this assumption are summarized in Table E-1. For example, deaths in the age interval $[x, x+1)$ occur, on average, at age $x+\frac{1}{3}$ if they occur in the lower triangle and at age $x+\frac{2}{3}$ in the upper triangle. Deaths in either triangle contribute, on average, $\frac{1}{3}$ of a person-year of exposure within the triangle where the death occurred. At the same time, all deaths result in an equivalent amount of lost exposure within their respective triangles (relative to what the individual would have contributed if s/he had exited the triangle as a survivor).

Table E-1. Implications of assuming uniform distribution of deaths within Lexis triangles (at age *x*)

	Lower triangle	Upper triangle
Average age at death	$x+\frac{1}{3}$	$x+\frac{2}{3}$
Average contribution (per death) to exposure within triangle	$\frac{1}{3}$	$\frac{1}{3}$
Average lost exposure (per death) within triangle	$\frac{1}{3}$	$\frac{1}{3}$

These relationships are not intuitively obvious and must be derived using calculus. The uniformity assumption implies that the probability density of deaths is 2 over the triangle (because the total area is $\frac{1}{2}$). The values in Table E-1 are then found by integrating over age and time. For example, the average age at death in a lower triangle is found by solving the following double integral:

$$\int_0^1 \int_0^t 2(x+s)\,ds\;dt = x + \frac{1}{3}\ . \qquad (E1)$$

Similarly, the average lost exposure for a death in the upper triangle is

$$\int_0^1 \int_t^1 2(1-s)\,ds\;dt = \frac{1}{3}\ , \qquad (E2)$$

and so forth.

Cohort death rates and probabilities

Death rates and probabilities of dying are simpler conceptually for cohorts than for periods. As depicted in Figure 10 (main text), cohort rates and probabilities are measured over the parallelogram that follows the lives of individuals who turn age *x* in one calendar year until their next birthday, at age $x+1$, in the following calendar year. We assume that there is no migration and deal with the case of a closed population, since the effects of migration on calculated rates and probabilities are negligible so long as migratory flows have the same direction and a similar magnitude over the interval.

For cohorts, the probability of dying between ages *x* and $x+1$ is found by first computing the probability of survival, p_x, which is the product of two components:

$$p_x = \frac{P}{P+D_L} \cdot \frac{P-D_U}{P} = \frac{P-D_U}{P+D_L}\ , \qquad (E3)$$

where $P = P(x,t+1)$, $D_L = D_L(x,t)$, and $D_U = D_U(x,t+1)$, as illustrated in Figure 10. Thus, the probability of surviving from age *x* to age $x+1$ is a product of the fraction surviving from age *x* to the end of the calendar year and the fraction surviving from the beginning of the next calendar year to age

$x+1$. It follows immediately that

$$q_x = 1 - p_x = \frac{D_L + D_U}{P + D_L} \qquad (E4)$$

is the cohort probability of dying in the age interval $[x, x+1)$.

To obtain the cohort death rate over this interval, we must first compute the exposure-to-risk. Consider the *P* individuals alive at the boundary between the two calendar years. If there were no deaths in the upper triangle, these *P* individuals would contribute a total of *P* person-years of exposure over the complete age interval. However, as shown above, the D_U deaths in the upper triangle result in an average lost exposure of $\frac{1}{3}$ person-years, which must be subtracted from *P*. On the other hand, the D_L deaths in the lower triangle contributed an average of $\frac{1}{3}$ person-years, which must be added to the total for the interval. Therefore, the person-years of exposure in this interval equal

$$E_x = P + \frac{1}{3}(D_L - D_U)\ , \qquad (E5)$$

and the cohort death rate is, by definition,

$$m_x = M_x = \frac{D_x}{E_x} = \frac{D_L + D_U}{P + \frac{1}{3}(D_L - D_U)}\ . \qquad (E6)$$

As reflected in this equation, for cohorts there is no difference, either conceptually or numerically, between the population death rate, M_x, and the life-table death rate, m_x.

Let a_x be the average number of years lived between ages *x* and $x+1$ by individuals who die in the interval. Under the assumption that deaths are distributed uniformly within Lexis triangles, we know that the average age at death is $x+\frac{1}{3}$ in the lower triangle and $x+\frac{2}{3}$ in the upper one (see Table E-1). It follows that

$$a_x = \frac{\frac{1}{3}D_L + \frac{2}{3}D_U}{D_L + D_U}\ . \qquad (E7)$$

We recall that, in classical life table notation, death rates, probabilities of dying, and average years lived in the age interval are related by the following formula:

$$q_x = \frac{m_x}{1 + (1 - a_x) \cdot m_x} \quad . \qquad (E8)$$

It is easy to confirm that these three quantities, when derived (as above) under the assumption of a uniform distribution of deaths within Lexis triangles, satisfy this equation. Thus, the three quantities are mutually consistent, even though they have been derived independently from death counts and population estimates. In the case of cohorts, this relationship is exact, and there is no need to give preference to either rates or probabilities in the calculation of cohort life tables.

Period death rates and probabilities

As depicted in Figure 9 (main text), period death rates and probabilities of death are measured over the 1x1 Lexis square that lies between exact ages x and $x+1$ during some calendar year. Therefore, these quantities reflect the blended experience of two birth cohorts. As explained in the main text, we begin by computing the period death rate within a 1x1 Lexis square, which we then convert to a probability of death by assuming a constant force of mortality over the age interval. Although this is a standard method, it lacks the desirable property enjoyed by our method for computing cohort quantities, since it is not reversible: if we begin by computing the probability of death directly from data and converting it to a death rate, we obtain a slightly different result. We adopt the method used here because it includes an explicit calculation of the exposure-to-risk, which is needed for statistical modeling, and because the link between death rates and probabilities of death is well defined.[31]

Our method for converting death rates into probabilities of death has already been described in the main text. Furthermore, it is a familiar technique (Preston *et al.*, 2001) and requires no particular

[31] It is common in the French demographic tradition to compute probabilities of dying directly from data, using the method of "partial quotients" (Pressat, 1980). In this tradition, death rates are also computed directly from data, based on an explicit calculation of the exposure-to-risk. The main difference, relative to the method used here, is that death rates and probabilities of death are not linked by an explicit formula. In practice, however, there is very little difference between empirical results obtained using the two methods.

justification here. Therefore, our only task in this section is to justify our method of computing the exposure-to-risk that forms the denominator of the period death rate. As before, we assume that there is no migration within the 1x1 Lexis square and deal with the case of a closed population, since the error introduced by this assumption is usually negligible.

To compute the exposure-to-risk in a 1x1 Lexis square, we are required to make an additional assumption. Consider the $N(x,t)$ individuals who attain exact age *x* and the $N(x+1,t)$ individuals who attain exact age *x*+1 in calendar year *t* (see Figure 9). Suppose, in each case, that the birthdays of these individuals (at age *x* and *x*+1) are distributed uniformly within the calendar year *t*. Then, neglecting deaths, the $N(x,t)$ individuals who attain exact age *x* in year *t* would have contributed, on average, $\frac{1}{2}$ person-year of exposure within the lower triangle; likewise, the $N(x+1,t)$ individuals who attain exact age *x+1* in year *t* would contribute, on average, $\frac{1}{2}$ person-year of exposure within the upper triangle. Thus, the major component of the exposure-to-risk in this case would be $\overline{N} = \frac{1}{2}\left[N(x,t) + N(x+1,t)\right]$. If the two distributions of birthdays (at age *x* and *x*+1) are not uniform within the calendar year but are nevertheless similar to each other, then the correct multiplier would differ only slightly from $\frac{1}{2}$. This type of mild non-uniformity can be safely ignored in computing the exposure-to-risk. However, larger departures from this uniformity assumption can be more problematic.[32]

[32] This assumption is violated most severely in situations where there are rapid changes in the size of successive cohort, owing to fluctuations in the birth series many years before. The worst situation is when a sharp discontinuity in births occurs in the middle of one calendar year, creating a cohort that is "heavy" at one end and "light" at the other. We have not attempted to correct our mortality estimates for the error introduced by such occurrences, which may result in artificially elevated or depressed levels of mortality along a diagonal of the Lexis diagram that follows the cohort(s) in question. The user should be aware of this possibility and not misinterpret the data.

In addition, the $D_L = D_L(x,t)$ deaths in the lower triangle result in an average lost exposure of $\frac{1}{3}$ person-years, which must be subtracted from $\overline{N}$. On the other hand, the $D_U = D_U(x,t)$ deaths in the upper triangle contributed an average of $\frac{1}{3}$ person-years, which must be added to the total for the interval. Therefore, the person-years of exposure in this interval can be estimated as

$$E_x = \overline{N} - \frac{1}{3}\left(D_L - D_U\right) . \qquad (E9)$$

Notice that $N(x,t)$ is equivalent to the $P(x,t+1)$ population aged *x* at the beginning of calendar year *t*+1 plus the D_L deaths in the lower triangle, and that $N(x+1,t)$ is equivalent to the $P(x,t)$ population aged *x* at the beginning of calendar year *t* minus the D_U deaths in the upper triangle. Therefore, we can substitute $\frac{1}{2}\left[P(x,t+1) + D_L + P(x,t) - D_U\right]$ for $\overline{N}$ in equation (E9). After simplifying, we get

$$E_x = \frac{1}{2}\left[P(x,t) + P(x,t+1)\right] + \frac{1}{6}\left(D_L - D_U\right) , \qquad (E10)$$

and thus the period death rate is

$$M_x = \frac{D_x}{E_x} = \frac{D_L + D_U}{\frac{1}{2}\left[P(x,t) + P(x,t+1)\right] + \frac{1}{6}\left(D_L - D_U\right)} . \qquad (E11)$$

Appendix F

Special methods used for selected populations

For the sake of comparability, we aim to follow the general principles described in this document for all populations included in the HMD. However, exact uniformity of methods is not always possible, because data at the required level of detail are not available in all situations. Therefore, in a few special cases, we have developed special methods to accommodate the realities of the available data. As of the current version of this document (dated May 31, 2007), the populations listed below have been treated with special methods. For the most up-to-date version of this table, go to the *Special Methods* link (http://www.mortality.org/Public/Docs/SpecialMethods.pdf) on the HMD website.

Population	**Special method**	**For more details:**
Belgium	Corrected counts of live births (1895-1923, 1919) and infant deaths (1886-1955, 1958-60) to include false stillbirths. Used special methods to accommodate missing deaths for 1914-18.	See Appendices 2 and 3 of the country-specific documentation
Bulgaria	Calculated population estimates for 1989-1992 by treating official estimates for 1988 as a "pseudo-census" and then applying the intercensal survival method.	See the "Data Quality Issues" section of the country-specific documentation
Canada	Adjusted death counts for missing information (e.g., sex, age, year of birth) and errors in the year of birth. In some cases, we aggregated the original death counts by Lexis triangle into 1x1 format because of apparent data quality problems.	See the "Death Count Data, Special Details" section of the country-specific documentation
England and Wales	Used estimates for the total population that incorporate the military population.	See the country-specific documentation
France	Corrected counts of live births and infant deaths (1899-1974) to include false stillbirths. Used estimates for the total population that incorporate the military population.	See the country-specific documentation
Germany	Corrected for overestimation of the old-age population for West German males in the most recent year.	See Appendix 2 of the country-specific documentation
Germany, East and West	Used special methods to estimate death, birth, and population counts separately for East and West Berlin since 2001.	See Appendix 2 of the country-specific documentation
Italy	Estimated the age distribution for death counts in 1893-94. Adjusted the census counts for 1871, 1921, and 1951 to cover the same territory as the death counts.	See Appendix 2 of the country-specific documentation

Population	Special method	For more details:
Lithuania	Used SR(85+) rather than SR(90+) to derive population estimates for almost extinct cohorts. The original death counts were aggregated for ages 99+.	See the "Population Count Data, Specific Details" and "Data Quality Issues" sections of the country-specific documentation
New Zealand	We adjusted the Māori and Non-Māori population counts for 1991-1995 to follow the older definition of ethnicity. Due to the fact that deaths in 1995 are classified by a mixture of the previous and current definitions of ethnicity, we introduced a special adjustment factor for Māori and Non-Māori deaths. In addition, we adjusted census counts for 1960 & 1970 to cover the *de jure* ("usual resident") population.	See Appendix 2 of the *Background & Documentation* file for the national population
Norway	Estimated the sex distribution for births during 1846-1915.	See the country-specific documentation
Portugal	Used SR(85+) rather than SR(90+) to derive population estimates for almost extinct cohorts.	See the "Population Count Data, Specific Details" section of the country-specific documentation
Russia	Prior to calculation of HMD estimates, i) the original death counts for 1959-1989 were aggregated for ages 99+, and ii) official population estimates in recent years were aggregated for ages 80+. Population estimates for almost extinct cohorts were derived using SR(80+) rather than SR(90+).	See Appendix 2 of the country-specific documentation
Spain	Corrected counts of live births and infant deaths (1930-1974) to include false stillbirths. Adjusted the census counts for 1940, 1950, 1960, & 1970 to cover the *de facto* population and the same territory as death counts.	See the country-specific documentation
Sweden	Adjusted death counts for 1863, 1865, 1868, & 1870 to match a secondary data source.	See the "Data Sources" section of the country-specific documentation
Switzerland	Adjusted deaths counts for females in 1878.	See the "Death Count Data, Deaths at 99+" section of the country-specific documentation
Ukraine	Prior to calculation of HMD estimates, the original death counts for 1959-1989 were aggregated for ages 99+.	See the "Data Quality Issues, Age Heaping in Deaths" section of the country-specific documentation
United States	Tabulated individual data to derive deaths by Lexis triangle (1959-2004). Adjusted population estimates to exclude the Armed Forces overseas (1940-1969) and the population of Alaska and Hawaii (1950-1958). Adjusted births for 1959 to include Hawaii. Used the extinct cohort method for ages 75+ during 1933-39 because the official population estimates extend only to age 75+.	See the country-specific documentation

References

Andreev, Kirill (1999). *Demographic Surfaces: Estimation, Assessment and Presentation, with Application to Danish Mortality, 1835-1995*. Ph.D. dissertation, Faculty of Health Sciences, University of Southern Denmark.

Andreev, Kirill (2001). "Kannisto-Thatcher database on population and death counts at older ages." Internal documentation. Rostock, Germany: Max Planck Institute for Demographic Research.

Coale, Ansley, and Paul Demeny, with Barbara Vaughn (1983). *Regional Model Life Tables and Stable Populations*. New York: Academic Press.

Keyfitz, Nathan (1985). *Applied Mathematical Demography*. New York: Springer-Verlag.

McNeil, Donald R., T. James Trussell, and John C. Turner (1977). "Spline interpolation of demographic data." *Demography* 14(2): 245-252.

Pressat, Roland (1980). *Demographic Analysis: Methods, Results, Applications*. New York: Aldine Publishing.

Preston, Samuel H., Patrick Heuveline, and Michel Guillot (2001). *Demography: Measuring and Modeling Population Processes*. Malden, Massachusetts, United States: Blackwell Publishing.

Thatcher, A. Roger, Väinö Kannisto, and James W. Vaupel (1998). *The Force of Mortality at Ages 80 to 120.* Odense, Denmark: Odense University Press.

Thatcher, A. Roger, Väinö Kannisto, and Kirill Andreev (2002). The survivor ratio method for estimating numbers at high ages. *Demographic Research* 6(1):2-15. Retrieved 4 April 2002 (http://www.demographic-research.org/).

Vallin, Jacques (1973). *La mortalité par génération en France, depuis 1899*. Paris: Presses Universitaires de France.