Human mortality beyond age 110

Jutta Gampe

Max Planck Institute for Demographic Research Konrad-Zuse-Str. 1, 18057 Rostock, Germany. E-Mail: gampe@demogr.mpg.de

Abstract. The International Database of Longevity (IDL) offers detailed information on thoroughly validated cases of supercentenarians. These data are used to estimate human mortality after age 110. The procedure properly accounts for the country-specific sampling frames in the IDL. The analysis confirms that human mortality after age 110 is flat at a level corresponding to an annual probability of death of 50%. No sex-specific differences in mortality could be found, and no time trend in supercentenarian mortality between earlier and later cohorts could be detected.

1 Introduction

The principal motivation for undertaking the effort of collecting data on supercentenarians in the International Database of Longevity (IDL; see Chapter 2 in this volume) is to estimate human mortality at the most advanced ages, based on information that is age-ascertainment bias-free and thoroughly validated. The shape of the hazard trajectory at the most advanced ages is interesting in itself, but it also has important implications for interpreting the general principles that rule human mortality.

While the exponential increase in the force of mortality, as described by the Gompertz distribution, is accepted for mid-adult and early old ages, there is general agreement that mortality increase slows down after about age 80 to 85, a phenomenon most likely to be explained by earlier selection of the frailer individuals in heterogeneous cohorts. Investigating how this slowing down continues into the highest ages has, however, so far been limited by the availability of sufficient high-quality data. The data provided by the contributors to the IDL now offer the opportunity to make our knowledge more complete.

H. Maier et al. (eds.), *Supercentenarians*, Demographic Research Monographs,
DOI 10.1007/978-3-642-11520-2_13,

Identifying (and subsequently validating) potential supercentenarians is a complex task that differs from country to country due to the different data sources available. The reports in Part II of this volume give a detailed account of the difficulties that needed to be resolved in each country. The sources available for identifying individuals in the relevant age-group has implications for the sampling frame, that is, for how individuals were selected for inclusion in the database. These sampling frames have to be taken into account in the estimating procedure to render valid inference. In particular, the way in which the individuals were identified implies certain truncation and censoring patterns.

The general patterns and the implications for the estimating procedure will be discussed in Section 2, followed by a description of the specific statistical model used. As censoring and truncation are particular forms of incomplete data, the expectation-maximization (EM) algorithm (Dempster et al., 1977) is a natural candidate for obtaining the actual maximum likelihood estimates. Section 3 explains the general principle, while the technical details are provided in the appendix at the end of the chapter. The results for supercentenarian mortality are reported in Section 4, and a summary of the findings concludes the chapter.

2 Sampling frames and likelihood

As the IDL involves individual data with ages given up to the day, we can model on a continuous age-scale. This is conceptually easier and allows for a leaner notation. We will therefore outline the general steps of the analysis in continuous time before moving on to a more specific model in the next section.

We are interested in the random variable X describing the distribution of human life spans after age 110. Its distribution can uniquely and equivalently be characterized by its density $f(x)$, its survival function $S(x) = P(X > x)$, or, most prominently in mortality analysis, its hazard $\mu(x) = f(x)/S(x)$. Once an appropriate distributional family is selected, indexed by an unknown parameter(vector) θ, this parameter will have to be estimated. Maximizing the likelihood function will be the method of choice, due to its good statistical properties overall.

Assuming independent individuals, the likelihood function $L(\theta)$ is the product of the individual contributions $L_i, i = 1, \ldots, n$, for the n individuals in the sample,

$$L(\theta) = \prod_{i=1}^{n} L_i \quad \text{or on log-scale} \quad \ln L(\theta) = \sum_{=1}^{n} \ln L_i. \tag{1}$$

The individual terms have to reflect exactly the information an individual observation contributes. This includes whether we have exact information on the age at death x_i, or whether the information is censored because we only know that the individual has survived a certain age (right-censoring), or has died between two ages (interval-censoring). The latter typically results if living supercentenarians are listed annually.

Truncated information has to be incorporated in the same way. An observation is truncated if the individual was selected into the sample only because he or she met a certain condition *related to the random variable investigated*. In our case, this means that an individual is included in the sample only because he or she survived a certain age (left-truncation), or because he or she died before reaching a particular age (right-truncation).

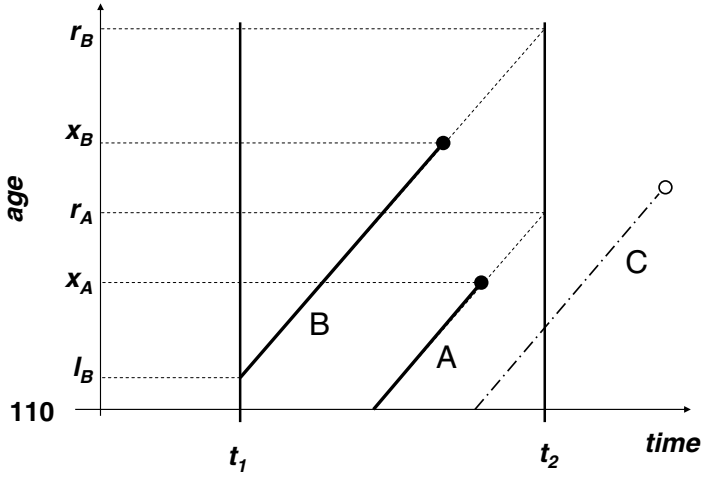

Fig. 1. Sampling frames: Right- and left-truncated observations.

There are several different sampling frames represented in the IDL data, but one of the most common is the identification of supercentenarians based on lists of deaths after age 110 that occurred between two

calendar times t_1 and t_2. Figure 1 shows individuals in such a sampling frame in a Lexis diagram.

Individual A, who turned 110 during the observation interval $[t_1, t_2]$, is observed only because he or she did not survive the age r_A that he or she would have reached at time t_2. This individual is right-truncated at age r_A. Assuming we have an exact age at death x_A of individual A, his or her likelihood contribution would have to be

$$L_A = P(X = x_A \,|\, X \leq r_A) = \frac{f(x_A)}{1 - S(r_A)}\,. \tag{2}$$

In the same way, individual B is right-truncated at age r_B. However, additionally, B is in the sample only because he or she survived his or her entry age l_B. If this individual had died before we would not have seen it in the sample. Thus, individual B is left-truncated at age l_B:

$$L_B = P(X = x_B \,|\, X > l_B, X \leq r_B) = \frac{f(x_B)}{S(l_B) - S(r_B)}\,. \tag{3}$$

In contrast, individual C, who crosses the observation interval but does not die in the interval $[t_1, t_2]$, is not seen in the sample.[1]

If we have age-at-death information only in an age interval, or survival information only, the respective numerators in (2) and (3) would have to be replaced accordingly. In general, truncation is concerned with the selective exposure pattern in the sample, while censoring deals with imprecise information on the event time. To practically estimate the parameters of interest, the model distribution, i.e., the specific form of $f(x)$, $S(x)$, and thereby $\mu(x)$, remains to be specified.

3 Statistical model and the EM-algorithm

The ultimate goal is to flexibly estimate mortality after age 110, without imposing a particular shape on the tail behavior of the distribution. As parametric continuous distributions determine the trajectory of the hazard in the limit, we have chosen a quasi-continuous approach (Pagano et al., 1994; Tu et al., 1993). The continuous distribution of X is approximated by a discrete distribution by clipping the age axis after

[1] For sampling frames that follow individuals from when they reach age 110 up until this so-defined cohort dies out, equation (3) still remains valid. $S(l)$ can be replaced by 1 (as in equation (2)), and the right-truncation condition can be pushed to infinity, i.e. $S(r = \infty) = 0$. The denominator in (3) then simply reduces to 1.

age $x_0 = 110$ into small intervals of length δ. The discrete probabilities p_j then correspond to

$$p_j = P(x_0 + (j-1)\,\delta < X \leq x_0 + j\,\delta), \qquad j = 1, \ldots, J.$$

The number of intervals J results from the requirement that the last interval covers the highest age at death. The discrete survival function correspondingly is denoted by $S_j = \sum_{k \geq j} p_k$, which leads to the discrete hazard $\mu_j = p_j / S_j$. If δ is chosen to be one year, then the μ_j directly give age-specific annual probabilities of death. Given that we have age-at-death information up to the day, we can choose δ much smaller than one year and thereby obtain a quasi-continuous estimate of $\mu(x)$ and $S(x)$. The unknown parameters θ in this model are the $J-1$ probabilities p_j (the last, p_J, automatically results from $\sum_j p_j = 1$).

If all individuals were fully observed, i.e., no censoring or truncation were present, then the log-likelihood (1) simply would be of multinomial form

$$\ln L(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} I_{ij} \ln p_j \,, \tag{4}$$

where $I_{ij} = 1$, if individual i dies in interval j, and zero otherwise. The maximum-likelihood estimates are $\hat{p}_j = \sum_i I_{ij}/n$.

The simplicity of the complete-data likelihood (4) makes an EM-algorithm an appealing solution. Starting from a current estimate $\hat{\theta}^{(m)}$, the I_{ij} are replaced by the expected numbers $E_{ij}^{(m)}$ that are to be seen, but that cannot be observed due to the truncation or censoring present, as in the case of individual C in Figure 1.[2] This is the E-step.

The M-step maximizes this pseudo-complete data likelihood to obtain the next estimate $\hat{\theta}^{(m+1)}$. The procedure is continued until convergence. The main step in the EM-algorithm is the calculation of the expected values $E_{ij}^{(m)}$. Details are given in the appendix.

4 Results for supercentenarian mortality

The strategy laid out in the previous section was applied to the data in the IDL as of October 31, 2008. Only the most reliable data in the IDL were included, i.e., data from countries with information that was assessed to be of validation level A (see Chapter 2). The cases included come from the countries listed in Table 1.

[2] Turnbull (1976) pictorially called these 'ghosts'.

Table 1. Number of supercentenarian cases included in the analysis.

Country	Cases	Country	Cases
Belgium	5	Nordic Countries	26
England & Wales	66	Quebec	10
France	49	Spain	28
Germany	17	Switzerland	4
Italy	37	USA	341
Japan	54		

The truncation and censoring patterns result from the sampling procedures in each country. Details on how individual cases were identified can be found in the country reports in Part II of this volume.

The total number of cases included was 637, of which 573 are females and 64 are males. Table 2 gives the ages at death in completed years. The birth cohorts of the individuals are listed in Table 3.

Table 2. Ages at death or at right-censoring of supercentenarians included in the analysis.

	110	111	112	113	114	115	117	119	122
Female	295	150	66	33	20	6	1	1	1
Male	29	17	10	4	3	1			

Table 3. Year of birth of supercentenarians included in the analysis.

1852–64	1865–69	1870–74	1875–79	1880–84	1885–89	1890–94	1895–99
6	10	47	85	152	248	80	9

As can be seen, about half (50.9%) of all individuals die within one year after becoming a supercentenarian, and about three-quarters (77.1%) die within two years after their 110^{th} birthday.

Figure 2, top panel, shows the estimated hazard for an interval-length $\delta = 0.25$, i.e., a quarter of a year. The results are given on the hazard scale, that is, the discrete probabilities p_j were transformed assuming a piece-wise constant hazard. The dashed line represents a hazard level of $-\ln 0.5 \approx 0.7$, which was obtained by Robine et al. (2005) based on a much smaller set of supercentenarians. This hazard

level corresponds to an annual probability of death of 0.5. As can be seen, the estimated hazard varies around this level, with stark fluctuations occurring after age 114 due to the small number of observations at these truly advanced ages. For the three most extreme observations, this model necessarily only gives three isolated spikes in the hazard. The corresponding log-survivorship curve is given in Figure 2, bottom. It clearly demonstrates the constant hazard up to about age 114 by its strikingly linear decline.

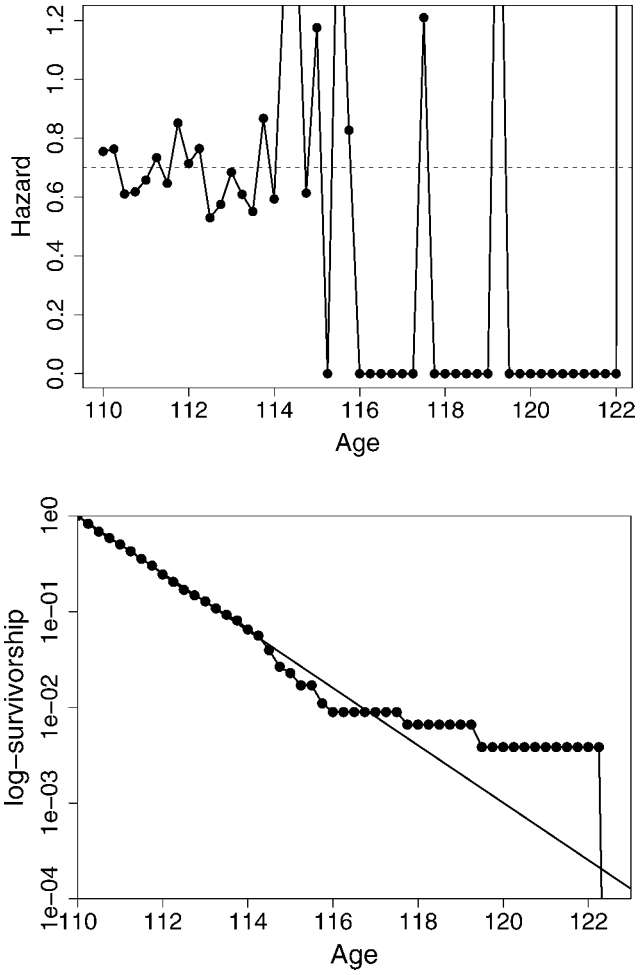

Fig. 2. Hazard (top) and log-survivorship (bottom) estimated for $n = 637$ supercentenarians.

Death rates at older ages have been declining in recent decades, and mortality improvement has been shifting into higher and higher ages. It is therefore natural to wonder whether a time trend in mortality can be observed in death rates after age 110.

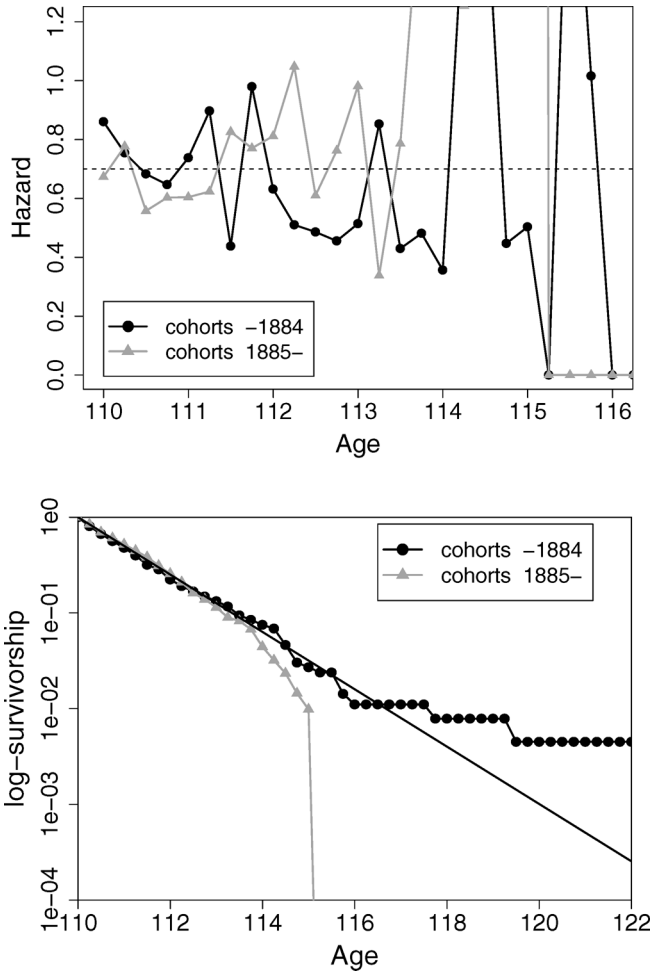

Fig. 3. Hazard (top) and log-survivorship (bottom) estimated separately for the earlier and the later cohorts in the dataset.

To check for such a time trend, we split the dataset into earlier and later cohorts. As we can see from Table 3, the number of observations rises sharply in the decade 1880-1890. To obtain a fairly balanced split,

the sample was divided into two groups: earlier cohorts, defined as birth cohorts up to 1884 (300 cases); and all individuals born 1885 or later (337 cases).

Figure 3 shows the hazard and the log-survivorship curves for early and late cohorts. No time trend in mortality of supercentenarians is supported by the result.

The large majority (almost 90%) of supercentenarians are women. While females enjoy lower mortality than males more or less at all ages, supercentenarians certainly are a highly selected group of individuals. Despite the comparatively small number of male cases, we estimated the hazard for male and female supercentenarians separately. The results are given in Figure 4, and show no significant sex-specific differences.

5 Conclusions

In summary, based on our analysis of 637 supecentenarian cases that were obtained in an age-ascertainment bias-free way, and that were thoroughly validated, we can state the following results.

- Human mortality after age 110 is flat at a constant level of $\lambda \approx 0.7$. This implies an annual probability of death of $q_x = 0.5$. This result confirms the previous analysis by Robine et al. (2005). Correspondingly, life expectancy after age 110 is about 1.4 years. Beyond the age of 114, data become too sparse to allow us to make reliable statements.
- No sex-specific differences can be detected. However, we have to be aware that only a small portion, about 10%, of supercentenarians are males, making the sample size for comparisons highly unbalanced.
- No differences in levels of mortality could be found between earlier and later cohorts.

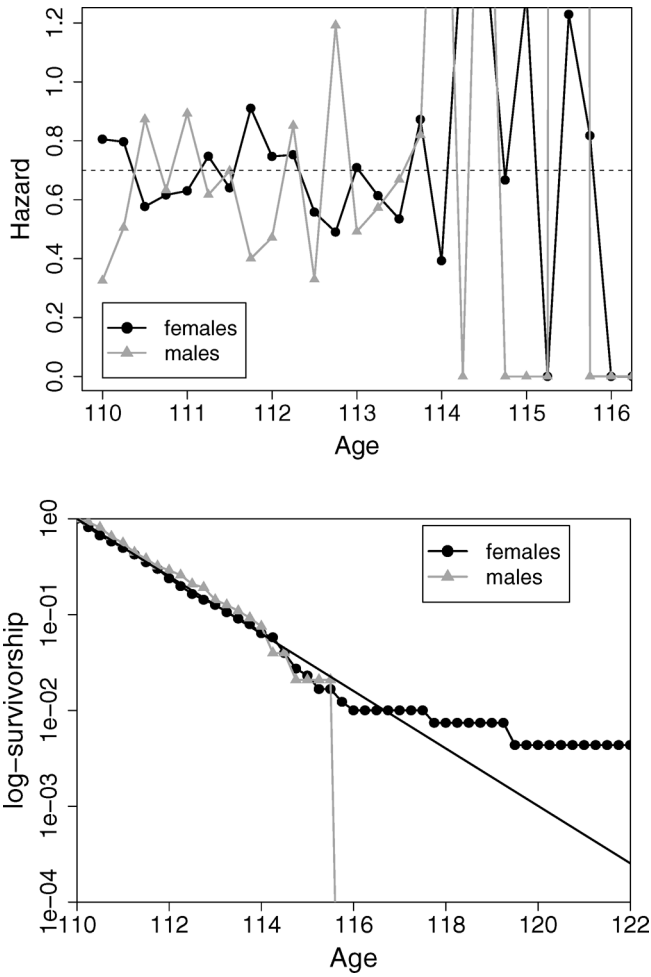

Fig. 4. Hazard (top) and log-survivorship (bottom) separately for male and female supercentenarians.

Appendix

For each individual i we define the censoring set $A_i \subset \{1, \ldots, J\}$ and a truncation set $B_i \subset \{1, \ldots, J\}$. If an exact age at death is observed, then A_i is a single number. If, for example, an individual is only known to have died between ages corresponding to the intervals j_1 to j_2, then $A_i = \{j_1, \ldots, j_2\}$. The same procedure is followed for the truncation sets B_i.

Similar to the I_{ij} in equation (4), we define $\xi_{ij} = 1$, if interval $j \in A_i$ and zero otherwise. And $\eta_{ij} = 1$ if $j \in B_i$ and zero otherwise. The expected values $E_{ij}^{(m)}$ that replace the I_{ij} in the likelihood (4) are the sum of two terms, which we denote by $c_{ij}^{(m)}$ and $d_{ij}^{(m)}$. The $c_{ij}^{(m)}$ are

$$c_{ij}^{(m)} = \xi_{ij} \frac{p_j^{(m)}}{\sum_{k=1}^{J} \xi_{ij} p_k^{(m)}}.$$

For exactly observed age l the $c_{il}^{(m)} = 1$ and $c_{ij}^{(m)} = 0$ for $j \neq l$. If the individual is censored the $c_{ij}^{(m)}$ give the expected value for interval j. Similarly, the d_{ij} give the expected number of individuals to be seen at j if not filtered by truncation:

$$d_{ij}^{(m)} = \frac{(1 - \eta_{ij})\, p_j^{(m)}}{\sum_{k=1}^{J} \eta_{ik}\, p_k^{(m)}}.$$

The $E_{ij}^{(m)}$ to be inserted into (4) are

$$E_{ij}^{(m)} = c_{ij}^{(m)} + d_{ij}^{(m)}.$$

Detailed derivations can be found in McLachlan and Krishnan (1997) or Pagano et al. (1994).

References

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the *EM* algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39: 1–38.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

Pagano, M., Tu, X. M., DeGruttola, V., and MaWhinney, S. (1994). Regression analysis of censored and truncated data: Estimating reporting delay distributions and AIDS incidence from surveillance data. *Biometrics*, 50: 1203–1214.

Robine, J.-M., Cournil, A., Gampe, J., and Vaupel, J. W. (2005). IDL, the international database on longevity. In *Living to 100 and beyond*, Living to 100 and Beyond Symposium, Orlando, FL. Society of Actuaries.

Tu, X. M., Meng, X.-L., and Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association*, 88: 26–36.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38:290–295.